HOW TO MEASURE IN EDUCATION

HOW TO MEASURE IN EDUCATION

WILLIAM A. McCALL, Ph.D.
ASSOCIATE PROFESSOR OF EDUCATION
AT TEACHERS COLLEGE, COLUMBIA UNIVERSITY

New York
THE MACMILLAN COMPANY
1927

Set up and electrotyped. Published January, 1922. Reprinted April, June, July, 1922; March, 1923; May, 1927.

PRINTED IN THE UNITED STATES OF AMERICA BY
THE BERWICK & SMITH CO.

PREFACE

The *science* of measurement is in its infancy and the *art* of measurement is younger still. Yet both have developed at such a phenomenal rate in the last few years as to make this movement for the mental measurement of children the most dramatic tendency in modern education. Educators returning from a few years' sojourn in foreign missionary fields testify that no recent change in educational practices compares with that effected by the growth of scientific educational and intelligence testing. During the war, and since, it was my privilege to confer with educational or military commissions sent here by several foreign governments to study our schools. These representatives testified that the extensive use of scientific mental measurement was one of the most distinctive features of American education. This book describes rather fully the meaning and methods of this movement.

The art of measurement is younger than the science of measurement because the abler workers have, of necessity, devoted their energies almost exclusively to the origination of foundational techniques. But the whole movement was so promising in the way of concrete assistance in meeting educational problems that practical educators have irresistibly demanded that the science of measurement be turned into the art of measurement almost overnight. Hence the last two or three years has witnessed a feverish effort to meet these demands. The result has been numerous mistakes and remarkable successes.

This book has a fourfold aim. First, it aims to present these successes and warn against a repetition of the mistakes.

Secondly, it aims to help forward the movement for mak-

ing teaching a genuine profession. Effective teaching requires a skill and a command of refined procedure in excess of that needed by the physician or surgeon. The medical profession is a genuine profession just because its members are experts in refined procedure. This book is one of the two which I have planned to show that tangible, learnable, refined techniques are possible in teaching and that even a "born" teacher will not, in a few years, be able to compete with a professionally *trained* teacher.

In the third place this book aims to meet the needs of the educators who are interested in both the How and the Why. Extremely elementary books in this field have done an admirable service in diffusing an interest in mental measurement. The more intelligent teachers are now asking, however, for a book which not only brings the art of measurement up to date, but which, in addition, goes sufficiently into the science of it that they may be able to use tests less blindly than heretofore.

The final aim of this book is to bring together in one convenient volume most of the techniques needed by those engaged in mental measurement. At present the worker in this field must go to one book to learn how to construct a mental test, to another book to learn how to give and use the results of the test, to another book to learn how to apply statistical methods, and to still another book to discover methods for graphic and tabular presentation. The expert will and should continue to consult these special treatises. Others do not like to give the necessary time. This book then, is really several small books in one. Furthermore, it has been so arranged that the reader can confine himself to the earlier and less technical portions or he can omit this and study the more technical chapters placed toward the end.

As I think through the contents of this book and reflect upon the influences which have made it, I am keenly conscious that it is both autobiography and biography. It is autobiography because several years of my own thought and

labor and much experimentation in the world's most stimulating educational laboratory, Teachers College and New York City, have gone into it.

This book is a biography because from cover to cover I see the tangible and relatively intangible evidences of my many teachers. My pupils have been my teachers. My contemporaries in mental measurement have been my teachers. In so far as my mention can accomplish it, I wish, however, to honor particularly five individuals. The first is, and has been, an elementary school teacher in the mountains of Kentucky and Tennessee. The second is President Emeritus of Cumberland College, Williamsburg, Ky. The third is President of Lincoln Memorial University, Cumberland Gap, Tennessee. The influence of the fourth is responsible for my interest in realms which transcend those represented in this book. The ideas of the fifth are so completely a part of me that it might be said that he wrote the book through an imperfect medium. They are Jesse Worley, E. E. Wood, George A. Hubbell, Frank M. McMurry, and Edward L. Thorndike.

WM. A. MCCALL

CONTENTS

PART ONE

How to Use Measurement

PART TWO

How to Construct and Standardize Tests

PART THREE

Tabular, Graphic, and Statistical Methods

APPENDIX

LIST OF TABLES

LIST OF DIAGRAMS

PART ONE

HOW TO USE MEASUREMENT

CHAPTER I. PLACE OF MEASUREMENT IN EDUCATION

CHAPTER II. MEASUREMENT IN CLASSIFYING PUPILS

CHAPTER III. MEASUREMENT IN DIAGNOSIS

CHAPTER IV. MEASUREMENT IN TEACHING

CHAPTER V. MEASUREMENT IN EVALUATING EFFICIENCY OF INSTRUCTION

CHAPTER VI. MEASUREMENT IN VOCATIONAL GUIDANCE

HOW TO MEASURE IN EDUCATION

CHAPTER I

PLACE OF MEASUREMENT IN EDUCATION

THESIS 1. "WHATEVER EXISTS AT ALL, EXISTS IN SOME AMOUNT" [1]

It is possible to become so immersed in the details of the measurement of pupil achievement as to lose sight of its fundamental significance. It is this absence of perspective which is responsible for two educational afflictions—the lop-sided enthusiast for the scientific measurement of education, and the equally unbalanced opponent of the movement. Educational measurement has a sort of philosophy. I have attempted to condense the main elements of this philosophy into a series of theses which may help those who have not had much time to think along these lines to appreciate the true place which measurement should have in education. The first of these theses is stated above.

Since all sane persons accept this thesis it needs no qualification, but a qualified thesis will suffice for our purpose, namely, whatever change the teacher makes in a pupil must be a change in an amount of something. We teachers will scarcely insist that our effort makes no change in amount. Even though such were the result of our effort it would not so much disprove the thesis but rather prove our own inefficiency.

There is an ever-dwindling group who strenuously oppose

[1] E. L. Thorndike, *The Seventeenth Year Book of the National Society for the Study of Education,* Part II, p. 16; Public School Publishing Co., Bloomington, Ill.

the practical implications of the above thesis. They claim to be interested in the emancipation of education from the quantitative idea. Their effort is directed toward the qualitative in education. According to them there is in every person a non-quantative quality—a

> " . . . something far more deeply interfused,
> Whose dwelling is the light of setting suns."

Did they truly "see into the life of things" they would realize that there is never a quantity which does not measure some quality, and never an existing quality that is non-quantitative. Even our halos vary in diameter.

THESIS 2. ANYTHING THAT EXISTS IN AMOUNT CAN BE MEASURED

At least half a dozen scales now exist by which it would have been possible to measure the quality of the Handwriting on the Wall. Faust said:

> "What she reveals not to thy mental sight
> Thou wilt not wrest from her with levers and with screws."

But science has enormously increased the subtlety of levers and screws, and our mental sight is obtuse compared to some of our present-day mental tests. Jesus tacitly accepted and practiced mental measurement when He estimated the quantity of faith on a mustard-seed scale.

It is possible to measure, at least crudely, an individual's love of a sunset or appreciation of opera. Theoretically the thesis is sound but whether practically we shall ever possess sufficient ingenuity to discover all the things that exist in amount and then measure them with any great accuracy, is a question. All that is necessary to accept for the present is that all the abilities and virtues for which education is consciously striving can be measured and be measured better than they ever have been. The measurement of initiative, judgment of relative values, leadership, appreciation of good literature and the like is entirely possible. We already have a scientific scale for the measure-

ment of poetic appreciation. The measurements may not be as exact as we might wish, but they would have value.

THESIS 3. MEASUREMENT IN EDUCATION IS IN GENERAL THE SAME AS MEASUREMENT IN THE PHYSICAL SCIENCES.

The two types of measurement are fundamentally alike because both measure physical manifestation. Neither adding ability, nor good intentions can be measured by plunging a thermometer into a pupil's spiritual medium, but they can be by measuring his behavior and judging his inner condition therefrom. Unless the witness is a habitual liar, psychologists can, with considerable success, determine by means of a breathing curve, when a witness is not telling the truth.

In a still invisible future it may be possible to secure a "movie" of a pupil's mental machinery when in operation and thus secure the desired information but for the present it is necessary to measure the product produced and, if desired, infer the inner condition of the pupil.

Measurement must frequently meet the objection of being too materialistic. Listen to Gilder in "The Poet's Protest."

> "O man with your rule and measure,
> Your tests and analyses!
> You may take your empty pleasure,
> May kill the pine, if you please,
> You may count the rings and the seasons,
> May hold the sap to the sun,
> You may guess at the ways and the reasons
> Till your little day is done."

To parody Wagner in "The Better Way," one would think that it was the purpose to measure human worth by the ell, the value of a life by the number of its years, the painter's canvas by the yard, or the work of the poet by the pound or bushel. A student writes: "Measurement should not be applied where spiritual factors and ideal values are involved." Those educators who protest most violently against any such measurement of the pupil are daily probing

his mental activity by methods which are comparable to the surgical operations of bygone ages. They find themselves in a position of disapproving the lover who estimates his lady's affection by the radius of the pupil of her eye under standardized lighting, and of approving the scientific father who soothes the mother for his punishment of their infant by saying: "I am not slapping an innocent soul but spanking a physiological reaction."

THESIS 4. ALL MEASUREMENTS IN THE PHYSICAL SCIENCES ARE NOT PERFECT

Physical measurements are, in general, more exact than educational measurements but education has no monopoly upon imperfect tests. There are tests which are now the rule in physical sciences for which an expert in educational measurements would blush. The general superiority of physical measurements is not due to the fact that they are radically different in kind. Physical measurements are subject to practically all the errors which trouble educational measurement. It is not that they do not exist in the former, but that they usually exist in such small amounts that the average person fails to see them. They are large enough to be the despair of experts in the various sciences. Thorndike [2] has given us an excellent statement of this point:

"Nobody need be disturbed at these unfavorable contrasts between measurements of educational products and measurements of mass, density, velocity, temperature, quantity of electricity, and the like. The zero of temperature was located only a few years ago, and the equality of the units of the temperature-scale rests upon rather intricate and subtle presuppositions. At least, I venture to assert that not one in four of, say, the judges of the Supreme Court, bishops of our churches, and governors of our states could tell clearly and adequately what these presuppositions are. Our measurements of educational products would not at

[2] Op. cit. p. 18.

present be entirely safe grounds on which to extol or condemn a system of teaching reading or arithmetic, but many of them are far superior to measurements whereby our courts of law decide that one trade-mark is an infringement on another."

But the imperfections of educational measurements are, in general, far more glaring than the majority of those made in physics, chemistry and like sciences. Some may have gotten the impression from certain emotional and quixotic radicals that standard tests are perfect instruments. This is far from the truth. They have numerous and decided limitations. The recent somewhat unbalanced, but doubtless necessary, propaganda is justified not because of the perfection of the tests recommended, but the much greater imperfection of the tests or lack of tests now in use.

A common criticism of educational measurement is that the tests measure a narrow, limited segment of a pupil's totality. Physical measurements tend to be more handicapped in this respect than educational measurements. Most of their measurements, such as measurements of length, width, weight, and temperature are exceedingly narrow abstractions and they are exceedingly useful too. A totality test for a pupil would certainly be useful but if we possessed one we would proceed immediately to construct tests for the detailed measurement of pupil abilities. Scales for the measurement of composition are useful, but scales for the measurement of the elements which go to make up composition are also useful. Teachers not only teach children "all over"; they teach them in detail. If tests are to aid instruction effectively, there is as much need for them to measure in detail as in totality.

THESIS 5. MEASUREMENT IS INDISPENSABLE TO THE GROWTH OF SCIENTIFIC EDUCATION

Exact measurement has made possible the rapid progress in the natural sciences. It has been stated that the amount

of soap used is an index of the civilization of a country. The exactness of measurement is a good index of the status of a science. Consider where science would be without its meter, gram, ampere, volt, ohm, watt, henry and the like. More than anything else it has been the absence of exact measurement which has kept education from the rank of a science. This plea for the development of those instruments which will make possible the progress of education as a science is made with knowledge of a recent statement by a prominent educator: "I think it would be disastrous if education were reduced to an exact science."

Richards,[3] in his presidential address before the American Association for the Advancement of Science, said: "Plato recognized, long ago, in an often-quoted epigram, that when weights and measures are left out, little remains of any art. Modern science echoes this dictum in its insistence on quantitative data; science becomes more scientific as it becomes more exactly quantitative."

In fact, measurement and education are like the twin girls whose hair the mother of many children braided together. Neither of the twins could move unless they both moved together.

Foote gives the above quotation in a nutshell when he says: "The day of guesswork must give way to definite facts supported by undebatable evidence."

There are those who tremble lest the development of education as a science will squeeze out of life its emotions and delicate perceptions. As well fear that woman suffrage or the "female" in industry will destroy gallantry among men. The roots of these fine things go too deep into human nature. In an especially unhappy mood Amiel writes: "Philosophy will clip an angel's wings," and again, "Science is a lucid madness engaged in tabulating its own necessary hallucinations." The basic function of Science is to help us to attain our objectives in the quickest and most economical way, whether the objectives be material or

[3] "The Problem of Radioactive Lead"; *Science*, Jan. 3, 1919.

spiritual. Science is frequently looked upon as materialistic chiefly because only those persons who seek material objectives have had the good sense to secure the aid of Science. Haeckel, who has just drawn the line under his life's work, must have had in mind the unnecessary inefficiency of idealism when he wrote: "No cosmic problem was ever solved or even advanced by that cerebral function we call emotion." For centuries education has been like an emotional dog chasing a frantic tail. We have had a long line of great educational thinkers from Plato through Pestalozzi, and Froebel . . . to Dewey and beyond. "The old order changeth, yielding place to new," but no one seems to know whether the old or the new is better. In fact, there is grave suspicion that we move in an orbit whose form is the circle. These educational leaders are not answering questions. They are asking questions which do not occur to others. They are proposing problems for experimentation. The final answer to every educational question, except one, must be left to the educational measurer and must await the development of education as a science.

THESIS 6. MEASUREMENT IN EDUCATION IS BROADER THAN EDUCATIONAL TESTS

This book is not entitled EDUCATIONAL TESTS because there are other methods of measuring pedagogical products. As previously stated, some estimate the quality of instruction by investigating the material equipment of libraries and laboratories and class rooms, or by the academic or professional training of the teacher. Others measure instruction by observing the teacher's method and by forming an opinion on the basis of these observations. Others base their judgments upon detailed observations of the behavior of pupils. Still others test the pupils by means of examinations. This book attempts to discuss the basic principles of measurement which apply not only to educational tests but to any sort of educational measurement. Some of the above methods of measuring educational results we are

likely to have with us for some time to come. In our zeal for improving tests proper, we should not neglect the refinement of these methods. The main emphasis should be, and in this book will be, upon tests, because they offer the best promise for exact measurement. For if we can trust the experience in other fields, measurement by means of some sort of instruments will gradually replace all other forms. Finally, the book is not entitled EDUCATIONAL MEASUREMENT because education is deeply concerned with measurements which are not exactly products of instruction but about which educators need to be critical.

THESIS 7. THERE ARE OTHER THINGS IN EDUCATION BESIDES MEASUREMENT

It will doubtless come as a surprise to the reader to hear one who is interested in the scientific measurement of education make such an admission. There are at least three other important factors in education, namely, pupil, methods and material, and goals. The teacher needs to know the psychology of the pupil, the proper goals to which the pupil is to be developed, and the methods and material which should be employed to develop the pupil from his initial ability to the desired goal. What does measurement have to do with this process?

THESIS 8. TO THE EXTENT THAT THE PUPIL'S INITIAL ABILITIES OR CAPABILITIES ARE UNMEASURABLE A KNOWLEDGE OF HIM IS IMPOSSIBLE

It has just been stated that a teacher needs the most intimate possible knowledge of a pupil in order to know what methods and materials to employ in order to help him most quickly to attain a desired goal. We partly know a pupil when we know the abilities and capabilities which he possesses. To determine the mere existence of an ability involves a crude measurement. But if we know no more than this we cannot tell whether a pupil has these abilities

in sufficient quantities to permit him to matriculate for a Ph.D. in the university or just enough to enter the kindergarten. We must know not only what qualities exist, but also in what amount they exist, and the more exactly we know this amount the better. Measurement is essential to a practical knowledge of psychology.

THESIS 9. "TO THE EXTENT THAT ANY GOAL OF EDUCATION IS INTANGIBLE IT IS WORTHLESS" [4]

We want to be able to answer at least three things about any goal: (1) What is the worth of the goal? (2) What is the location of the goal? (3) Is the pupil moving toward or from the goal? Measurement is necessary to answer each one of these absolutely vital questions. Suppose it be said that one goal of instruction is to produce in the pupil an ability to write. The worth of this goal depends upon an exact or crude measurement of how much penmanship contributes to the efficiency of a number of other superior activities. The goal has advanced little beyond perfect intangibility until it is located. How much ability to write? What speed? What quality? Even the worth of the goal cannot be answered until this location is made, since the worth varies with the quantity. The very words *how much* imply and in fact require measurement. Finally, it is necessary to answer the question: Is the pupil moving toward or from the goal? Without measurement the question is unanswerable.

THESIS 10. THE WORTH OF THE METHODS AND MATERIALS OF INSTRUCTION IS UNKNOWN UNTIL THEIR EFFECT IS MEASURED

The purpose of certain methods and materials is to help the pupil grow toward a certain goal. Do the methods employed accomplish their purpose? We cannot tell without employing measurement. For aught we know, the methods may be actually vicious. They may be forming

[4] I am indebted to F. M. McMurry for this thesis.

habits which not only do not lead toward the goal, but which may be building up difficulties for another method by a subsequent teacher. It is equally true that the comparative worth of different methods and materials is unknown until their effect upon the pupil is measurable. This means that measurement is indispensable to the experimental selection of the most economical educational conditions.

Thus, measurement is everywhere in education and in our daily lives. Measurement is no rare freak. It gets up with us in the morning and goes to bed with us at night. The mile stone, the hand of the watch, the humble cup in the kitchen, the lengthening shadows of the trees on the grass, the spacing of the year into seasons, all indicate how ubiquitous measurement is. And measurement is just as immanent in the whole educational process as in life in general. There are other things in education besides measurement but they have no value so long as they are dissociated from it.

THESIS II. MEASUREMENT OF ACHIEVEMENT SHOULD PRECEDE SUPERVISION OF TEACHING METHOD

Education is now being measured in two ways. When a child, I watched two coal miners lift a derailed car. Their efforts illustrate these two methods of measurement. A lever and fulcrum were brought, but the lever broke. A stronger lever was secured, but the fulcrum was too far from the car. Finally the proper adjustments were made and the car was lifted. Whether or not the car was lifted could be determined in two ways; (1) by measuring the length of the lever, the resistance of the fulcrum and the ground under the fulcrum, the weight of the men, the point of application of their weight, the distance of this point to the fulcrum, the distance from the fulcrum to the car, the weight of the car; or (2) simply by determining whether the car was actually lifted.

It is a fair assumption that the crucial purpose of elemen-

tary education is to make certain changes in children. To this end we have surrounded them with levers and fulcra in the shape of books, pictures, maps, tools, playthings, pedagogical methods and with teachers who will utilize these instruments as leverages to produce the desired changes.

Again it is a fair assumption that the schools should know whether their levers, fulcra, etc., are really producing the changes desired. As in the case of the derailed car, there are two methods of measuring these changes; (1) strength of lever, length of leverage, etc., become the number and nature of the books in the libraries, map facilities, blackboard space, and such, and the weight of the men becomes the number of diplomas possessed by the teacher or else the amount of her skill in making provision for motive, initiative and such on the part of her pupils. (2) Whether the car is actually lifted is comparable to measuring directly the changes in the pupils.

Doubtless our relatively primitive ancestors held conferences to discuss the advisability of such and such arrangements of lever and fulcrum in lifting a weight. Of course such possible discussions never were and never could be settled until the crucial measurement—the *direct* measurement was made. It would be of inestimable value to know whether the presence of certain books in the schoolroom, or the possession of a certain amount of professional training on the part of the teacher and the like are prerequisites of certain defined changes in pupils. Without such *ancillary* measurements by teachers and supervisors, the conditions for pupils' growth cannot be arranged in advance with certainty. But we shall not arrive at such knowledge except through direct measurement. We certainly cannot claim to know the exact casual relation between defined changes in pupils, and most of the paraphernalia with which the pupil is now surrounded. In spite of our ignorance of these causal relations, the chief method of supervision at present is to attempt to judge the presence or absence or amount of presence of these levers and fulcra.

THESIS 12. MEASUREMENT IS NO RECENT EDUCATIONAL FAD

Judging from the vituperation that has been heaped upon it, and the efforts that have been necessary to propagate it, one would think that scientific measurement was something absolutely novel. As a matter of fact, educators are, and have always been confirmed users of measurement—measurement of a kind. For several generations teachers have been employing tests which, to the uninitiated observer, would differ from standard tests in only one respect. The teacher's test is usually written on the blackboard while the standard test is usually printed on paper. Present a standard test to a teacher or principal who never heard of one, and neither will recognize that it is possessed of any peculiar virtues or patent dangers. Ayres tells us: "If Dr. Rice is to be called the inventor of educational measurement, Professor E. L. Thorndike should be called the father of the movement." And yet, if the great majority of us had thought of standard tests before Dr. Rice, or scaled tests before Dr. Thorndike, we probably should not have deemed the ideas worth enough to spend time upon or dangerous enough to frantically bury.

The writer's experience with the critics of standard tests convinces him that these critics have but two important objections, first, tests are not available for measuring all the aims of instruction, and, second, tests are sometimes misused. The first objection calls for, not the disuse of tests, but greater zeal in the extension of tests. The second objection calls for zeal, not against tests, but against their misuse. The closest students of scientific measurement are rarely its opponents and at the same time they are its severest critics. They are the severest critics because their criticisms are pertinent and because they are aware of numerous defects invisible to the casual observer.

Measurement in education did not suddenly leap into existence. It has had a gradual evolution, or rather it has been on a plateau for centuries. A student's theme informs

us that: "Educational measurement is ancient as a fact, medieval as a process and modern as a science. Half of Solomon's proverbs are tests for wisdom." The Chinese had a far-flung system of testing which was a sort of beginning for the Hillegas Composition Scale. The Roman father considered his son's literary education finished when his son could read the Roman Law from the tablet in the public forum. Little progress was made beyond the conventional, formal examination until 1894. Rice conceived the idea of a comparative test to be used in measuring the results of instruction in many schools. Out of the comparative test grew norms, for the use of a comparative test upon many schools yields norms. It was the genius of Thorndike that made possible the next advance. Utilizing the Cattell-Fullerton equal-distance theorem, he devised a scale unit for the measurement of educational achievement. This marks the beginning of scientific educational measurement. Stone's Arithmetic Tests worked out under the direction of Thorndike and published in 1908, represent a sort of transition from the Rice comparative tests to the Thorndike Handwriting Scale published in 1909. Subsequent students of Thorndike's have elaborated the statistical technique for the construction of educational scales. Hillegas, Buckingham, Trabue, and Woody constructed respectively the Composition Scale, Spelling Scale, Language Scale and Fundamentals of Arithmetic Scale.

The movement for the scientific measurement of education has spread with great rapidity. Courtis has been particularly successful in disseminating an interest in tests. Hence it is appropriate that he should have directed the testing in the first formal survey where tests were employed. The survey was the New York City Survey of 1911-12 and the tests used were the Courtis Arithmetic Tests. Since that time the movement has grown apace until now tests and scales are in daily use throughout this country and around the world. Every school survey relies upon tests as one of its chief instruments for evaluating the efficiency of the

schools being surveyed. The Gary Survey by the Rockefeller Foundation spent over $10,000 upon tests of pupils. Most of the universities and many of the colleges and normal schools give courses in educational measurement. Several universities have a bureau of research which coöperates with the school communities in its region in the measuring of education. There are now about twenty-five formal city bureaus of research and the number is increasing. Great foundations like the Rockefeller Foundation are forwarding this movement. All are familiar with the splendid work of Ayres in connection with the Russell Sage Foundation. Hundreds of isolated workers are adding impetus to the movement through their earnest study of educational problems by means of scientific measurement. But perhaps the greatest single force for the advancement of scientific educational and psychological measurement will prove to be the war work of the Psychology Committee of the National Research Council. Yerkes,[5] chairman of this committee, writes:

"It is already evident that the contributions to methods of practical mental measurement made by this committee of the National Research Council, and by the psychological personnel of the army, are profoundly influencing not only psychologists, but educators, masters of industry and the experts in diverse professions. New points of view, interest and expectations abound. The service of psychological examining in the army has conspicuously advanced mental engineering, and has assured the immediate application of methods of mental rating to the problems of classification and assignment in our educational institutions and our industries."

In the words of an enthusiastic but beginning student, "the importance of this measuremental (!) movement has been realized in education."

[5] Robert M. Yerkes, "Report of the Psychology Committee of the National Research Council"; *The Psychological Review*, March, 1919.

THESIS 13. TESTS WILL NOT MECHANIZE EDUCATION OR EDUCATORS

There seems to be a feeling that tests favor the so-called mechanical or conservative rather than radical methods in education. When properly used, they favor neither one. Ultimately tests will be the judge to give an impartial decision as to which method is the more effective. Until scientific measurement is extended, however, no decision between the two methods can be reached, because present tests cannot measure some of the most important aims of both educational conservatives and radicals. Suffice it to state here that present standard tests when improperly used may easily cause a greater mechanization of education, but when properly used they may easily be the salvation of education from too great a mechanization. The defense of this statement will appear later.

Much less is there any ground for believing that tests will squeeze the humanity out of teachers. The teacher should be the master of the instrument, not *vice versa.* It is to be hoped that there are no teachers like the farmer who was uncertain whether he was working to support ten cows or they were working to support him. If there are such teachers, tests cannot injure them. They are beyond injury. It is not probable that because a teacher can measure a pupil's ability in handwriting her interest in the finer things of life will evaporate. There is food for reflection in the statement by Thorndike that it is not the mothers who weigh their babies least often who love them most.

THESIS 14. TESTS WILL NOT PRODUCE A DEADLY UNIFORMITY

Perhaps it would be more accurate to say: tests NEED not produce a deadly uniformity. Tests need not destroy individuality in pupils. There exists in the minds of some a fear that composition, handwriting and drawing scales placed in the schoolroom before the pupils or even used by

the teacher to measure pupil product, will tend to discourage individuality. This fear is justified if there are, in a school, teachers who will instruct pupils to write their compositions just like the compositions on the scale, or to make their penmanship look just like the penmanship on the handwriting scale. My faith in the common sense of the members of the teaching profession is so high that I do not hold it important to argue this question further. It must be evident to anyone that composition and handwriting scales are measuring instruments and not models to be imitated, except in so far as the increasing quality of the specimens on the scale are goals toward which to strive. In fact, when such scales are properly used, they should increase individuality. By placing before the pupils definite objectives, the scales tend to increase interest in attaining these objectives, and it is a truism of psychology that the most prolific source of varied products and hence originality and individuality, is a powerful interest. Let the destiny of a nation depend upon the development of anti-submarine devices, or the favor of a king depend upon the production of an original literary masterpiece, or the issue of a baseball championship contest depend upon the development of a new strategy, and the conditions are very favorable for individual initiative. Thus scales are favorable to individuality when they are used as measuring instruments and for the location of definite objectives. They were never intended for models to be imitated. The teacher should urge the pupil to write a composition which is of as high a quality as a certain scale specimen and not to write a composition which is just like that specimen. Trabue once remarked that it is possible to feed an infant according to its weight without feeding it with the scoop.

CHAPTER II

MEASUREMENT IN CLASSIFYING PUPILS

I. Classification by Intelligence Tests

Bases and Objectives of Classification.—There are three main types and several minor types of measurement which may be used as bases of classification. The three chief types are intelligence measurements, educational measurements, and pedagogical measurements or teachers' marks. Medical measurements are frequently used to classify together pupils who are anemic. Chronological measurements have been used at Fairhope to classify pupils into life groups. There will be considered here only the three main types as they are used to classify pupils, not by separate subjects but by a sort of average of all subjects. The technique of classifying by separate subjects will prove easy to the one who masters the technique to be described.

The first fundamental objective of classification is to *put together those of equal educational status.* It is believed that homogeneous groups will make more satisfactory progress, due to the fact that the teacher can teach such a group almost as one pupil. The needs of all pupils are then closely similar. The work can be more exactly adapted to all. It saves the wear and tear on the teacher of continually shifting adjustment from one grade of ability to another. Franzen has described the instruction of teachers in non-homogeneous groups thus, "they mystify the lower quarter and bore the upper quarter."

The second fundamental objective of classification is to *put together those who will progress at equal rate.* At the best, periodic reclassification will be necessary. These will

need to be much more frequent if provision is made for equal initial ability only and not for equal rate of progress. Provision should be made for both. To make perfect provision for equal rate of progress would require a knowledge of each pupil's interest, industry, physiological limit, etc. Fortunately a simpler method is available which will prove sufficiently accurate for practical purposes.

Classifying by single subjects rather than a cross section of all subjects does help some but not greatly. Any one subject in the elementary school is divided into a multitude of subordinate mental traits which may or may not be psychologically akin. There is as much difference between different parts of geography as between geography and history. The correlation between ability in addition and ability in subtraction is not much closer than the correlation between ability in addition and ability in grammar.

The above are the legitimate objectives of classification together with the justifications for these objectives. But there are illegitimate objectives to be guarded against. It is difficult to improve upon Judd's [1] summary of them.

"Sometimes the school allows a pupil to move up a grade or class, although it is known that he has not done the work below, because the parents of the child have influence and it does not seem safe to antagonize them.

"Sometimes the pressure of numbers in the lower grades or classes is so great that the teacher sends a pupil on in order to make room for the younger pupils, even when it is evident that the pupil will not be able to carry the higher work.

"Sometimes the teacher in a given grade is anxious to unload the backward or disorderly and therefore incompetent pupil on someone else, and since the open road is into the next higher grade, the child is sent on.

"Promotion is sometimes controlled by the calendar. Because the date for closing the schools has arrived, and the

[1] Charles H. Judd, *Introduction to the Scientific Study of Education*, pp. 109-110; Ginn and Company.

long vacation is at hand, pupils are declared to have completed the work whether they have or not.

"Sometimes it is more or less explicitly argued that the backward pupil is larger than the other children of like intellectual attainments and he should therefore be sent to the upper-grade room where the seats are larger."

Relation Between Mental Age and Quality of School Work.—The close relation between mental age and quality of school work is now never questioned. When any pupil fails to make satisfactory progress in his school work, the first step toward finding an explanation is usually to get a measure of the pupil's mental age. There is a substantial correlation between a teacher's mental-age rating for her pupils and her marks upon their school work. This is to be expected, however, for the excellence of a pupil's school work is the chief means the teacher has to estimate the pupil's mental age. But the same close relation is also shown by objective tests of intelligence. Terman reports a correlation of .725 between mental age and the quality of work in the first grade. McCall [2] reports a correlation of .78 in the sixth grade. Dickson [3] studied five first-grade classes and found that only 2 of the 33 retarded children had normal mentality. He concluded that, while there may be contributory causes, low mentality is undoubtedly the chief cause of the retardation of 31 of these 33 children.

Not chronological age, physical size, and a variety of other criteria, but ability to do the work is the real criterion for classification. Terman, Dickson, Whipple, and others have shown that a pupil's mental age is an excellent index of the quality of work a pupil will be able to do. In his study at Urbana, Whipple shows that mental tests give a more accurate classification than teachers' judgments and school marks. It therefore seems an inevitable conclusion that

[2] Wm. A. McCall, *Correlation of Some Psychological and Educational Measurements;* Bureau of Publications, Teachers College, N. Y. C., 1916.

[3] See Lewis M. Terman, *The Intelligence of School Children,* p. 64; Houghton Mifflin Company, 1919.

classification by mental age or its equivalent is superior to classification by teachers' judgments. Hence, Terman suggests that intelligence tests be given to all pupils, and that each pupil be placed in a grade closely corresponding to his mental age.

Relation Between Mental Age and Present Grade Position.—The best estimate is that 25 per cent of the pupils in any grade belong mentally in a lower grade and 25 per cent in a higher grade. In sum, there is both too much acceleration and too much retardation—too much acceleration of the stupid and too much retardation of the intelligent. Data collected by Terman [4] shows not only that almost every grade contains pupils with mental ages ranging from eight to fourteen but also that errors in classification bear more heavily upon bright pupils than dull pupils.

Still further evidence of the extent to which the bright pupil is penalized is found in Table 1. Strayer finds that the total chronological over-ageness is 33.5%. The other two columns show that this per cent is not exaggerated. When this 33.5% over-ageness is contrasted with 4.3% under-ageness some conception may be gotten of the injustice being done to young pupils with high mental ages. Late entrance, absences, and the fact that the school is adjusted to a higher than 100 I.Q. child explains part of this difference between over-ageness and under-ageness. Due to these causes, the per cent of under-ageness should not be expected to equal the per cent of over-ageness, but the two per cents should approximate each other. There are as many pupils whose mental age is above normal as below normal. Dickson has shown that the chief cause of over-ageness is a mental age below normal. Hence a mental age above normal should mean a corresponding under-ageness. But no investigation has revealed anything like a corresponding under-ageness.

[4] Lewis M. Terman, *The Intelligence of School Children*, p. 26; Houghton Mifflin Co., 1919.

TABLE I

Retardation and Acceleration in Towns and Cities (Adapted from Strayer, Morton, and Salt Lake City Survey Report).

Amount of Retardation or Acceleration	Strayer[5] 318 Cities	96 Cities and Towns of Nebraska[6]	Salt Lake City
Over-age 1 year	19%	16.3%	26.7%
Over-age 2 years	9.5%	7.6%	11.2%
Over-age 3 years	3.8%	3.3%	3.7%
Over-age 4 yrs. or more	1.3%	1.4%	1.2%
Total over-age	33.5%	28.6%	43%
Total under-age	4.3%		

Since the process of classifying by mental age and I.Q. is similar to the process of classifying by educational age and E.Q. the former may be discussed with the latter.

II. Classification by Educational Tests

Measurement of a Sample School.—Recently I was asked to bring my class and measure on one day all the pupils in a small school. Educational tests were to be used.

The first step in carrying out this project was to select the tests to be used. The tests selected were Thorndike's Reading Scale Alpha 2 both Parts I and II, Thorndike's Visual Vocabulary Scale A-2x, Trabue-Kelley's Language Completion Scale, ten words from each of six different columns of Ayres' Spelling Scale, Woody's Addition, Subtraction, Multiplication, and Division Scales, Series B, and Composition scored by the Nassau Extension of the Hillegas Scale. No test was used which could not be given to any pupil in the entire school, and which did not measure an important phase of the school's work.

When tests are to be used for reclassifying a school the

[5] George D. Strayer, *Age and Grade Census of Schools and Colleges, Bulletin 451,* p. 144; U. S. Bureau of Education, 1911.

[6] W. H. S. Morton, "Retardation in Nebraska"; *Psychological Clinic,* Dec., 1912, and Jan. 19, 1913.

beginner will do well to select tests according to the following special principles:

1. The test should be uniform for all grades being reclassified. Some tests have one form for, say, grades III, IV, and V and another form for grades VI, VII, and VIII. Unless the scores are comparable from one form to the next, and they rarely are, such tests will cause difficulty.

2. The test should yield a single score. Some tests yield both speed and accuracy scores. Such tests serve a useful diagnostic purpose, but the beginner is likely to experience difficulty in attempting to employ such tests for reclassification.

3. The test should measure an important phase of the school's work. Unless the school is to be classified by subjects, the different tests should measure different subjects as a rule.

The second step was to select and train examiners. They were trained by having them actually apply under observation the particular test assigned to them.

The third step was to apply the tests according to standard procedure. To prevent a collision between examiners the plan shown below was devised.

As an examiner was detailed to a class the number of the period under the appropriate grade was circled. Just as soon as an examiner finished his test he returned to the central office, reported his completion and a cross was drawn inside the circle. Immediately afterward the examiner whose turn was next was sent to the unoccupied class. A chart like this shows the director at a glance, what classes are and are not occupied and just whose turn is next. Observe that the testing periods are numbered from 1 to 7 in the vertical column under Grade III. This is because there are more tests than classes. Had there been seven classes and only six tests, the testing periods would have been numbered from 1 to 7 horizontally opposite Test I.

In accordance with this plan every pupil in the school above Grade II was tested.

		III	IV	V	VI	VII	VIII
Reading	Test I	1	2	3	4	5	6
Completion	Test II	2	3	4	5	6	7
Add. and Sub.	Test III	3	4	5	6	7	1
Composition	Test IV	4	5	6	7	1	2
Mul. and Div.	Test V	5	6	7	1	2	3
Vocabulary	Test VI	6	7	1	2	3	4
Spelling	Test VII	7	1	2	3	4	5

The fourth step was to score the tests and to compute pupil scores.

The fifth step was to tabulate pupil scores. The scores are shown in Table 2. The detailed tabulation by test elements, which was sent to the teachers, is not shown.

The sixth step was to compute the median (later chapter) score on each test for each grade. These are shown in Table 2 just above the grade medians for the preceding year.

The seventh step was to tabulate norms for each test and grade. These are shown in Table 2. A few of the tests were standardized for mid-year. But our tests were given at the end of the year. Hence before any fair comparisons could be made it was necessary to alter the norms to fit the end of the year. The following shows an approximate method for converting the mid-year norms for a test into June norms.

	III	IV	V	VI	VII	VIII
Mid-year norm	10	14	18	21	23	24
June norm	12	16	19.5	22	23.5	24.5

Computation of Composite Scores.—The next step was to compute a composite score for each pupil. As shown in Table 2, the first pupil, *Ant.*, made the following scores on the nine tests: 3, 35, 58, 31, 11, 0, 3, 3, 2.8. A composite of these scores could be made by the simple process of summing them. The sum of these scores is 146.8. The composite score as computed by us, however, is 91. To sum scores just as they stand is to give the score on the spelling test twice as much influence as the score on the reading test, and the score on the vocabulary test thirty

TABLE 2

The Scores Made on Nine Tests by 98 Pupils in Grades III Through VIII of a Certain School, Together With the Age, Composite of All Tests, Educational Age, Educational Quotient (E.Q.), and Reclassification for Each Pupil, the Medians for Two Successive Years, and the Norms for Each and All Tests for Each Grade, and the Q_1, Q_3, Q, and Multiplier for Each Test.

Mos. Age	Gr. III	Rea.	Comple.	Voc.	Spel.	Add.	Sub.	Mul.	Div.	Comp.	Com-pos.	Ed. Age.	E.Q.	Cl.
109	Ant.	3	35	58	31	11	0	3	3	2.8	91	122	112	5
111	Bau.	5	13	21	13	5	0	4	5		54	109	98	4
131	Dor.	12	25	26	14	7	0	0	3	3.8	76	117	89	4
92	Kim.	9	30	56	27	7	0	7	5	2.8	92	123	134	5
103	Lon.	2	13	20	8	9	1	3	3	1.1	44	105	102	4
112	Mil.	10	24	27	13	9	0	4	3	2.8	73	116	104	4
130	Pie.	7	17	29	10	6	0	1	0	2.8	54	109	84	4
139	Rob.	9	18	12	5	11	0	4	6	3.8	71	115	83	4
139	Soh.	0	6	14	2	6	0	0	3	1.9	28	99	71	3
154	Sie.	3	5	13	1	10	6	4	3	1.1	39	103	67	4
131	Sim.	11	16	19	8	6	6	4	1	1.9	60	111	85	4
122	Sch.	10	15	29	41	7	0	3	5	3.8	79	118	97	4
168	Wek.	3	8	39	0	6	6	5	2	1.1	44	105	63	4
149	Whi.	4	20	19	15	6	4	5	3	2.8	65	113	76	4
104	Wal.	4	11	20	2	5	0	3	0	1.9	38	103	99	4
133	Whit.	3	13	19	13	7	0	0	1	1.1	38	103	77	4
104	Wad.	3	6	7	2	5	0	1	3	0	20	96	92	3
119	War. R.	13	33	81	31	10	0	3	4	3.8	108	128	108	5
147	War. J.	19	27	97	31	3	10	7	2	3.8	116	132	90	5
Med. 1919		5.5	16.5	21.5	13.5	7.1	0.7	3.9	3.4	2.5	58	111	90	
Med. 1918														
Norm		8.0	18.5	30.0	19.6	9.0	6.0	3.5	3.0	2.1	71	115	100	

Mos. Age	Gr. IV	Rea.	Comple.	Voc.	Spel.	Add.	Sub.	Mul.	Div.	Comp.	Com-pos.	Ed. Age.	E.Q.	Cl.
130	Mil.	9	28	47	19	12	10	7	5	3.8	105	127	98	5
146	And.	8	16	26	7	14	7	4	5	1.9	70	115	79	5
147	Ada.	10	21	58	13	13	9	9	6	2.8	98	125	85	5
119	Bla.	10		80	34	9	4	9	6	2.8	109	129	109	5
124	Die.	13	26	46	22	15	8	10	5	3.8	112	130	105	5

(Data for Grade IV continued on next page)

TABLE 2—continued

Mos. Age	Gr. IV (cont.)	Rea.	Comple.	Voc.	Spel.	Add.	Sub.	Mul.	Div.	Comp.	Com-pos.	Ed. Age.	E.Q.	Cl.
124	Erh.	10	29	44	20	7	0	4	5	2.8	85	120	97	5
143	Law.	15	30	60	23	11	4	2	3	3.8	104	127	89	5
127	Lan.	8	29	34	33	12	4	9	4	2.8	98	125	98	5
139	Mye.	21	48	119	44	13	9	14	9	3.8	172	158	114	7
143	Mei.	9	24	36	17	10	0	5	2	2.8	77	117	82	5
136	Peh.	18	18	82	36	13	5	4	1	2.8	100	125	92	5
125	Per.	5	36	14	12	6	3	0	0	1.9	66	113	90	4
111	Rea.	18	30	69	33	12	7	11	7	2.8	124	135	122	6
136	Sin.	11	40	52	16	13	5	4	4	2.8	116	132	97	5
136	Van.	22	39	108	34	14	10	10	7	3.8	154	149	110	7
130	Tow.	7	24	32	21	9	9	10	8	2.8	94	123	95	5
160	Tra.	12	23	51	12	12	8	6	3	2.8	92	123	77	5
Med.	1919	10.8	29.0	51.5	21.5	12.6	7.3	7.5	5.4	3.0	90	125	97	
Med.	1918	12.0	29.3	53.5	36.6	9.2	7.4	6.7	3.7	2.8	105	127		
Norm		15.0	25.0	65.0	30.4	11.0	8.0	7.0	5.0	2.6	107	128	100	
Mos. Age	Gr. V													
129	Amb.	15	38	96	29	13	8	10	5	5.0	143	144	112	6
129	Buc.	18	36	84	26	10	8	7	5	2.8	124	135	115	6
139	Cat.	17	40	104	31	15	10	9	4	2.8	140	142	102	6
132	Fer.	18	48	104	37	10	8	8	3	2.8	142	143	108	6
132	Hoy.	21	43	104	43	13	8	8	8	5.0	151	148	112	7
136	Key.	16	38	78	32	13	6	10	4	3.8	133	139	102	6
133	Lin.	19	30	60	28	10	0	0	4	5.0	109	129	97	6
163	Lig.	16	42	68	34	10	7	3	4	5.0	132	139	85	6
189	Mil.	17	36	88	29	13	8	8	4	3.8	133	139	74	6
135	Sim.	24	41	57	38	13	9	12	7	3.8	149	147	109	6
123	Sco.	23	39	113	43	10	11	11	9	5.0	165	154	125	7
142	Sny.	9	25	47	2	9	9	6	4	1.9	82	119	84	5
145	Sch.	17	34	78	21	11	9	7	4	3.8	124	135	93	6
133	Tra.	19	36	73	28	12	9	11	4	5.0	140	142	107	6
Med.	1919	18.0	38.5	81.5	30.5	12.	8.8	8.7	4.9.	4.1	138	141	105	
Med.	1918	17.7	37.3	82.0	75.7	12.	10.2	9.3	7.5	3.8	155	149		
Norm		20.0	30.5	83.0	37.8	14.	10.0	11.0	7.0	3.0	137	141	100	

TABLE 2—continued

Mos. Age	Gr. VI	Rea.	Comple.	Voc.	Spel.	Add.	Sub.	Mul.	Div.	Comp.	Com-pos.	Ed. Age.	E.Q.	Cl.
135	Ant.	25	53	109	51	14	8	13	10	3.8	181	163	121	8
142	Fra.	27	60	126	55	18	14	17	9	5.0	213	191	135	10
149	Hol.	20	40	93	29	14	10	11	7	5.0	155	149	100	7
163	Ham.	11	27	50	11	13	7	7	5	2.8	98	125	77	5
138	Kim.	26	67	117	54	15	8	10	8	5.0	200	181	131	9
156	Lin.	22	39	90	44	15	10	13	7	5.0	164	154	99	7
157	Lan.	21	49	111	59	16	12	14	11	5.0	190	170	108	8
148	Mid.	16	38	85	37	18	11	12	8	5.0	157	150	101	7
144	Mit.	23	45	115	57	15	10	12	8	5.0	180	162	113	8
146	Ort.	13	42	104	49	13	9	10	4	3.8	147	146	100	7
150	Pat.	24	44	120	50	12	9	5	3	2.8	151	148	99	7
150	Pug.	23	48	114	52	14	13	11	11	6.0	185	165	110	8
183	See.	9	26	39	10	9	8	4	3	1.9	80	118	64	5
157	Smi.	17	39	91	46	12	9	7	3	3.8	137	141	90	7
147	Spi.	24	53	119	57	12	11	10	10	6.0	193	173	118	8
144	Tay.	23	39	70	44	10	8	7	4	6.0	150	147	102	7
157	Woo.	16	33	59	32	11	10	6	2	2.8	115	131	83	6
151	Erh.	24	42	114	45	13	9	9	10	5.0	170	157	104	7
Med.	1919	23.	42.5	106.5	47.5	14.	10.0	10.7	8.	4.7	169	152	102	
Med.	1918	19.	35.0	85.3	58.7	12.	10.5	9.8	7.	3.8	152	148		
Norm		24.	38.0	95.0	47.7	16.	12.0	15.0	10.	3.6	165	154	100	
Mos. Age	**Gr. VII**													
184	Bel.	24	56	124	59	15	12	12	11	5.0	200	181	98	9
184	Cum.	30	39	36	17	12	8	6	5	2.8	127	137	74	6
167	Dam.	23	49	122	59	17	12	15	10	6.0	200	181	108	9
188	Hen.	25	57	98	41	16	12	15	12	6.0	201	182	97	9
138	Hal.	11	46	49	8	7	5	4	3	2.8	103	127	92	6
171	Kas.	24	41	95	38	13	11	15	7	3.8	162	153	89	7
192	Lut.	24	59	112	45	17	11	14	10	6.0	202	182	95	9
169	Lan.	24	49	118	56	15	12	14	11	5.0	193	173	102	8
164	Hal.	27	45	100	30	15	11	11	11	3.8	169	156	95	8
187	Heb.	25	71	124	57	14	14	12	10	3.8	209	188	101	10
157	Lon.	21	44	105	48	13	6	8	4	3.8	152	148	94	7
163	Pug.	29	67	127	58	19	11	16	13	6.0	229	205	126	11

(*Data for Grade VII continued on next page*)

TABLE 2—continued

Mos. Age	Gr. VII (*cont.*)	Rea.	Comple.	Voc.	Spel.	Add.	Sub.	Mul.	Div.	Comp.	Com-pos.	Ed. Age.	E.Q.	Cl.
149	Pil.	27	47	108	41	14	12	12	6	5.0	179	162	109	8
177	Rus.	28	47	97	53	18	12	12	8	6.0	192	172	97	8
173	Soh.	27	48	110	48	18	13	14	8	5.0	191	171	99	8
169	Sch.	20	46	86	20	13	10	8	3	3.8	143	144	85	7
167	Scb.	23	48	113	49	15	15	16	10	6.0	196	176	105	9
165	Sny.	20	40	88	42	17	13	15	8	6.0	175	160	97	8
171	Woo.	23	49	118	59	15	12	14	11	5.0	193	173	101	8
Med. 1919		24.6	48.3	108.5	48.3	15.5	12.2	14.1	10.1	5.0	188	172	97	
Med. 1918		23.5	46.5	91.0	47.5	13.7	10.6	11.1	9.3	5.3	175	160		
Norm		28.0	40.0	108.0	50.3	18.0	13.0	17.0	13.0	4.1	188	167	100	
Mos. Age	**Gr. VIII**													
171	Car.	25	57	108	47	16	13	10	9	6.0	198	179	105	9
187	Cat.	26	55	119	58	12	10	10	4	6.0	190	169	90	9
199	Dor.	28	49	97	54	18	12	11	6	6.0	191	171	86	9
158	Deg.	28	57	120	53	18	14	18	10	6.0	217	194	123	10
172	Lan.	25	45	104	58	14	13	15	9	7.2	197	178	103	9
204	Lig.	25	45	109	38	12	10	7	8	5.0	167	155	76	8
182	Roa.	24	51	98	29	15	12	12	10	6.0	184	164	90	8
179	Ric.	27	57	123	57	15	12	13	10	6.0	208	187	104	9
169	Wal.	25	66	120	55	18	14	17	12	7.2	230	206	122	11
197	Wid.	32		119	60	18	13	17	13	6.0	232	208	106	11
184	Woy.	18	39	95	57	14	9	8	5	7.2	167	155	84	8
Med. 1919		25.9	55.5	109.5	55.5	15.8	12.8	12.5	9.8	6.2	204	178	103	
Med. 1918		28.8	55.3	117.0	75.7	15.8	12.3	14.3	10.8	6.0	216	193		
Norm		30.0	43.5	117.0	54.4	18.5	14.5	18.0	14.0	4.8	199	180	100	
	Q_1	10.7	27.0	47.5	17.8	10.3	5.2	5.1	3.9	2.9				
	Q_3	24.5	48.6	109.7	49.2	15.2	11.4	12.5	9.1	5.3				
	Q	6.9	10.6	31.1	15.7	2.5	3.1	3.7	2.6	1.2				
Multipliers		1	1	1/5	1/3	1	1	1	1	5				

times as much weight as the score on the composition test. Competent judges are substantially agreed, however, that reading is certainly not less important than spelling, and vocabulary is not thirty times as important as composition. It is a very common practice, however, to sum scores just as they appear without giving any thought to this matter of weighting.

Tests should be weighted according to the variability of their scores. They should not, as is frequently supposed, be weighted according to the size of their scores. Which of these two tests exercises the most influence upon the composite?

Pupil	Test I	Test II	Composite
a	6	403	409
b	10	404	414
c	2	403	405
d	1	402	403

Test I has more weight than Test II because the variability of the scores in Test I is greater than the variability of the scores in Test II. The scores in Test I range from 1 to 10; the scores in Test II range only from 402 to 404, or, what is equivalent, from 2 to 4. If we multiply all the scores of Test II by 5, the range will be increased, and will then range from 2010 to 2020, i. e., 10 points. If we divide the scores of Test I by 5, or what is equivalent multiply them by 1/5, there will be a corresponding shrinkage of their variability.

The actual process of computing the composite for the scores of Table 2 was as follows:

1. The scores on the reading test of all the pupils in the entire school were thrown into a frequency distribution. (Later chapter.) Similar distributions were made for each of the other tests.

2. The Q_1, Q_3, and Q were computed for each test. (Later chapter.) They are shown at the end of the table. The Q_1 and Q_3 serve no other purpose than to yield Q, which is a measure of variability. The Q's show that, if the scores

were summed just as they stand, the vocabulary test with its Q of 31.1 would have greatest weight; and composition with its Q of 1.2 would have least weight.

3. The multipliers shown below the Q's were selected so as to re-weight the tests in rough accordance with my idea of how the tests should be weighted. The following shows the Q's, the multipliers and the new weight or new Q given.

Tests	I	II	III	IV	V	VI	VII	VIII	IX
Q	6.9	10.6	31.1	15.7	2.5	3.1	3.7	2.6	1.2
Multiplier.	1	1	1/5	1/3	1	1	1	1	5
New Q ..	6.9	10.6	6.2	5.2	2.5	3.1	3.7	2.6	6

The completion test or Test II was given most weight, not because it was considered more significant than the reading tests, but because it was a saving of labor to leave its Q unchanged. This completion test is a reliable and generally excellent test. It is one of the best intelligence tests and it was desired that the composite be a good index of intelligence. It was preferred to run the risk of giving it too much weight than of giving it too little, especially when leaving the Q unchanged reduced labor. Reading and completion could have been given identical weight by dividing 10.6 by 1.5. But guesses at the weighting which tests should receive will necessarily be so inaccurate that it is foolish to increase one's labor by using as multipliers or divisors other than whole numbers. Reading or Test I received the next largest weight. Vocabulary or Test III was given slightly less weight than reading, and this is probably as it should be. Composition or Test IX was given about the same weight as reading and vocabulary. Spelling or Test IV, being only an element of composition, was given slightly less weight than Composition. The tests of arithmetic fundamentals or Tests V, VI, VII, and VIII, were given individually the least weight of all. This was not because arithmetic is unimportant in school work, but because these tests were four in number and even then measured only a small section of the work in arithmetic. Since there were

four tests, arithmetic actually received a total weight of 11.9 or (2.5 + 3.1 + 3.7 + 2.6) as against 23.7 or (6.9 + 10.6 + 6.2) for the tests which may be classed as reading tests. These reading tests were given a greater combined weight than the arithmetic tests because reading is prerequisite to more of the total work of the school than the fundamentals of arithmetic. As many of these questions of weighting as possible should be settled by the refined, yet laborious, technique of partial correlation and regression equations.

4. All the pupil scores, grade scores, and norms for each test, were multiplied by the multipliers selected for that test. This means that nothing at all was done to the reading scores since their multiplier was 1. The same was true for completion. The vocabulary scores were divided by 5; the spelling scores were divided by 3; the arithmetic scores remained unchanged; and the composition scores were multiplied by 5. All these products and quotients do not appear in Table 2. In the original computation they were written between the columns of the table in red ink.

5. The weighted scores were summed to get a composite score for each individual, a grade composite, and a norm composite. The following illustrates the fourth and fifth steps for pupil *Ant*.

Test	I	II	III	IV	V	VI	VII	VIII	IX	Composite
Ant.	3	35	58	31	11	0	3	3	2.8	
Multiplier	1	1	1/5	1/3	1	1	1	1	5	
Weighted Score..	3	35	12	10	11	0	3	3	14	91

Transmutation of Grade Norms Into Age Norms.—Table 2 shows the educational age and E.Q. not only for each pupil in School X but also for the medians and norms of each grade. How in detail were these computed? It was impossible to compute the educational age of a pupil on any test until age norms were determined for the tests. Unfortunately it is the custom to report grade norms but not age norms for educational tests. No age norms being avail-

able, it became necessary to transmute grade norms into age norms. The third-grade norm on the spelling test reported in Table 2 is shown to be 19.6. Suppose it is known that the average chronological age of all third-grade pupils is 9 years. Knowledge of this would permit the conversion of the grade norm into an age norm. It could be said that the norm in spelling for average nine-year-olds is 19.6. In similar fashion all the grade norms could be converted into age norms.

The above conclusion may not be exactly true. Just because the median age of all third-grade pupils is nine years and the median score in spelling is 19.6, it does not necessarily follow that if all the nine-year-old pupils scattered through several grades were tested their median score would be exactly 19.6. There is, however, every reason to believe that it would be approximately 19.6. This method of transmuting grade norms into age norms assumes that the two would be identical. The method is a temporary one. It should be discarded as soon as properly determined age norms are available.

A rather thorough search did not reveal any widespread study which gives for the United States the average or median chronological age of the pupils in each grade. Enough data has been found, however, to permit the computation with a fair degree of accuracy of the average age of the pupils in each grade. Ayres [7] gives a frequency distribution showing the age of entering the first grade of 13,868 pupils who were about to graduate from the eighth grade in 29 cities. It has been computed from this table that the median age of entering first grade is 80 months. It is barely possible that pupils who remain to graduate from the elementary school tend to enter earlier or later than children in general. Any such difference if it exists at all is probably slight. Hence it is assumed that the median age at which pupils enter the first grade is 80 months.

[7] Leonard P. Ayres, *The Relation Between Entering Age and Subsequent Progress Among School Children, Bulletin No. 112;* Russell Sage Foundation, N. Y. C.

At least three studies show how long it takes the average pupil to complete each grade. In the above study Ayres found that the average time for the average pupil to complete each grade was 12.8 months (including vacation). Terman [8] found the average grade interval in terms of mental age to be 12.6 months. While acting as statistician for the Psychology Committee of the National Research Council in the preparation of the National Intelligence Tests, Kelley determined that the average time required for the average pupil to pass from one grade to the next was 13.2 months. It has been concluded, therefore, that the average time required for the average pupil to pass from grade to grade is roughly 13 months.

Having determined the age when the average pupil enters the first grade, and the average number of months required by him to pass from grade to grade it was possible to construct Table 3 for computing educational ages. The second and third columns of Table 3 may be used for any set of tests. The first column will depend upon the particular tests selected.

TABLE 3

The Norm Composite for, and Average Age in May of Pupils in Each School Grade. A Table for Converting Pupil Composites Into Educational Ages.

Norm Composite	Average Age	Grade
0 (est)	89	I
35 (est)	102	II
71	115	III
107	128	IV
137	141	V
165	154	VI
188	167	VII
199	180	VIII
216 (est)	193	IX
230 (est)	206	X
246 (est)	219	XI

[8] Lewis M. Terman, *The Intelligence of School Children,* p. 94; Houghton Mifflin Co.

The above table should be interpreted viz: Reading from right to left, pupils in Grade I have, in May, an average chronological age of (80 + 9) or 89 months, and the norm composite for Grade I is estimated to be zero. The average age for Grade II is (89 + 13) or 102 months with an estimated composite of 35. The average age for Grade III is (102 + 13) or 115 months with, not an estimated but, a known composite of 71. The average age of each succeeding grade has been determined by adding 13 months to the average age of the preceding grade. All the composites beyond the eighth grade are estimated.

Grade I is given an average age of 89 months instead of 80 months because the norms for all the tests shown in Table 2 are either May norms or have been computed forward to May. The average pupil in the third grade, say, was 115 months old when the norms for these tests were actually or arbitrarily determined. The interval between Grades VIII and IX (first year high school) is probably nearer 14 than 13 months, but in the absence of exact information it was preferred to keep constant the increment of 13 months.

Grade I was assigned a norm composite of zero because the average first grade pupil would probably make a zero score on these tests. It was estimated that the norm composite for Grade II was roughly half-way between zero and the norm composite for Grade III which is 71. The norm composites of 71, 107, etc., through 199 are the last numbers under each grade in the column headed *Composite* in Table 2. Each of these numbers is the weighted composite of the norms for the grade in question for all the tests. Any common sense method might be used to estimate the norm composites for grades beyond the eighth. The increase of Grade VII over VI and of Grade VIII over VII were averaged, the resulting 17 was added to the norm composite of Grade VIII, namely, 199. The composite of Grade X was found by averaging the increase of Grade IX over VIII and Grade VIII over VII and adding the resulting 14 to 216

and so on. It was necessary to thus extend the table below Grade III and above Grade VIII because there are pupils in Table 2 who have educational ages below Grade III and above Grade VIII.

The above table was designed to convert each pupil's *composite* score into an educational age. It can be used just as well to convert each pupil's score on each test separately into an educational age or subject age for that test. To do this it is necessary to substitute the norm score for each grade for the test in question in the place of the norm composite. If we were constructing an educational age table for the reading test in Table 2, for example, the norm scores of 8, 15, 20, 24, 28, and 30 would appear in the first column of Table 3 beside Grades III, IV, V, VI, VII, and VIII respectively. Appropriate norms could be estimated for lower and higher grades. In this way it would be possible to compute a reading age and Reading Quotient for each pupil in reading and classify the school for reading only. By constructing, in similar fashion, as many tables as there are tests in Table 2 it would be possible to compute for each pupil as many educational ages and E.Q.'s as there are tests, and thus classify by subjects if this is desired.

Again, a median of each pupil's educational ages and E.Q.'s on the nine tests would give a final composite measure of his educational age and E.Q. respectively. Or again it would be possible to determine the median of his nine educational ages and divide this final median once for all by his chronological age to get his final E.Q. If desired, the educational ages could be weighted according to the significance of the tests from which they were derived just as the original scores in Table 2 were weighted in computing the composite for each pupil. It was decided, instead, to compute each pupil's educational age and E.Q. just once and that through the composite of his original weighted scores.

Computation of Educational Age and E.Q.—The actual process of computing educational age and E.Q. was as follows: The first pupil in Table 2, namely, pupil Ant.,

has a composite score of 91. According to Table 3, had his composite score been 71 he would be given an educational age of 115, for he would have an educational status equal to the average 115-months-old child. Had his composite been 107 he would be entitled to an educational age of 128 months. Since his composite was 91 his educational age is between 115 and 128 months. Interpolating, it is found to be $115 + \left[\frac{128 - 115}{107 - 71} \times (91 - 71)\right] = 115 + \left(\frac{13}{36} \times 20\right) = 122$. Thus pupil Ant. has an educational age of 122 months. His chronological age as shown in Table 2 is 109 months. Since $\text{E.Q.} = \frac{\text{educational age}}{\text{chronological age,}}$ pupil Ant. has an E.Q. of $\frac{122}{109}$ or 112. Both his educational age and E.Q. are recorded in the last two columns but one of Table 2. Pupil War J. has an educational age of 132 and an E.Q. of 90, computed viz.: $\text{Ed. Age} = 128 + \left[\frac{141 - 128}{137 - 107} \times (116 - 107)\right] = 128 + \left(\frac{13}{30} \times 9\right) = 132$ months. $\text{E.Q.} = \frac{132}{147} = 90$. In this way an educational age and E. Q. were computed for every pupil and for the norm for each grade.

The 1919 third-grade median educational age could be computed in the same way, or could be found by taking the median of the educational ages of the third-grade pupils. The latter method was used and yielded a median of 111. Either method gives approximately the same result. The E.Q. for the 1919 medians could be computed either by dividing 111 by the median chronological age of the pupils or by taking the median of the E.Q.'s of the third-grade pupils. For reasons which will not be discussed here the two results will not be exactly identical. The E.Q. was found by taking the median of the pupils' E.Q.'s. The educational age and E.Q. for the norm must necessarily be

as shown in Table 2 because the computation of all pupil educational ages and E.Q.'s assumes that the norm third-grade educational age and E.Q. are 115 and 100 respectively.

In actually computing pupil educational ages a very much finer table than Table 3 was used. The working table showed the norm composite corresponding to every month. This saves the annoyance of interpolating, because with such a table no further calculation is necessary. Educational ages are read directly. Compare the following Grade III-IV portion of the working table, for example, with this same portion of Table 3.

Norm Composite	Average Age	Grade
71	115	III
73.8	116	
76.6	117	
79.3	118	
82.1	119	
84.9	120	
87.6	121	
90.4	122	
93.2	123	
95.9	124	
98.7	125	
101.5	126	
104.2	127	
107.0	128	IV

The age interval from Grade III to Grade IV is 13 months. The composite interval is 36 points. If 13 months equals 36 points, then one month is equivalent to 2.77 points. Hence the table begins with 71 and increases by 2.77 for each additional month until 107 is reached. Other grade intervals were spaced off the same way.

Educational Age *vs.* Mental Age.— Educational age when determined by a proper team of educational tests is probably superior to mental age for realizing the first objective of classification, namely, *to bring together pupils of*

equal educational status. Educational age is superior to mental age for this purpose because it and it alone reveals directly what pupils are of equal status educationally. Educational age measures this directly. Mental age measures educational status only indirectly. It has already been shown that there is a close relation between mental age and true educational status, but there are many forces operating to prevent this correlation from being perfect. A pupil's educational status is a resultant not only of his mental age but also of his health, attendance, attitude toward school work, industry, etc. Educational age takes into account both mental age and all these other factors which condition the quality of school work. Mental age, as usually tested, reveals the effect of these other factors but to a less extent.

Again, educational age is superior because it prevents the pupil from skipping valuable portions of the curriculum. If the curriculum has been properly constructed most of what is ahead is not likely to be so valuable as an equal amount of what is behind.

Finally, educational age is superior because it prevents the skipping of pre-requisite portions of ability hierarchies. Work in the elementary school is of a rather hierarchical nature. Even geography and history have certain prerequisites only a short distance below them. This point should not be stressed too much because gifted pupils have a phenomenal capacity to fill up really vital gaps. But educational age, particularly when it rests upon educational tests for the more continuous subjects, does guarantee that the pupil will not be handicapped by large gaps in his abilities.

Franzen has demonstrated, in the case of pupils whose educational age is markedly below mental age, that by specially promoting them and by otherwise applying educational pressure the educational age could be made to approximate the mental age within one year. It would be interesting to learn whether this progress could not have been

secured just as well, if not better, by keeping them at all times in the grade or grades closest to their educational age and applying the pressure there.

Mental age is, however, superior to educational age for classifying pupils in the primary grades. In the present stage in the evolution of educational tests and prognostic tests for special abilities, mental age is probably the best basis of classification for high school and college freshmen also, though some schools follow the practice of determining classification on the basis of educational tests of the progress made during the first week or weeks of school.

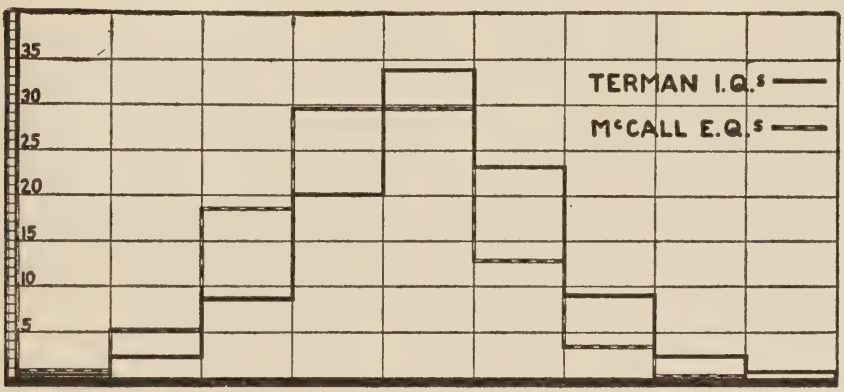
E.Q. *vs.* I.Q.—The second fundamental objective of classification is *to bring together pupils who will progress at equal rate.* Probably the best way to prophesy what the rate of progress will be is to find out what the rate of progress has been. Educational age and mental age considered apart from chronological age tell us almost nothing about the past rate of progress. Rate of progress is shown by E.Q. and I.Q. If the pupil's intelligence has developed rapidly his I.Q. is proportionally high—above 100; if it has developed slowly his I.Q. is low—below 100. Similarly, if his educational ability has developed rapidly his E.Q. will be proportionally above 100, and if it has developed slowly his E.Q. will be proportionally below 100.

Like educational age and mental age, E.Q. and I.Q. compare more easily than they contrast. It is their close similarity that strikes the student first. Fig. 1 brings out the similarity of their distribution. The solid line shows Terman's [9] distribution of I.Q.'s for unselected pupils. The dotted line blocks out the frequency surface of the distribution of E.Q.'s of about 500 pupils in a certain New York City school. (This school will be referred to hereafter as School Y.) The form of the E.Q. distribution closely approximates that of the I.Q. distribution. The I.Q.'s center at 100 while the E.Q.'s center five or so points below 100. This tendency for the E.Q.'s of School Y to be below

[9] L. M. Terman, *The Measurement of Intelligence,* p. 66; Houghton Mifflin Co.

the I.Q.'s for children in general is not surprising. The school is below norm on the educational tests, which is partly explained no doubt by the fact that on the average the children have a heredity which all observers were agreed is intellectually below par. Two diagrams comparing the E.Q. with the I.Q. for these same children would be almost identical.

A more detailed study of the data from School Y revealed a constant tendency for low I.Q.'s to be below the corre-



	56-65	66-75	76-85	86-95	96-105	106-115	116-125	126-135	136-145
McCALL –	0.66%	5.2%	18.6%	29.6%	29.6%	12.8%	3.4%	0.22%	-
TERMAN –	0.33%	2.3%	8.6%	20.1%	33.9%	23.1%	9.0%	2.3%	0.55%

FIG. 1. Comparison of E.Q.'s for 500 Pupils Tested by McCall with the I.Q.'s of Unselected Pupils Tested by Terman.

sponding E.Q.'s and for high I.Q.'s to be above the corresponding E.Q.'s. Several possible explanations of this phenomenon, apart from the method and error of computation, may be suggested. One is that stupid pupils, always in danger of failing, receive from the teacher the individual attention which not only belongs to them, but also that which belongs to the gifted pupils. This would raise the E.Q. of the stupid to the detriment of the E.Q. of the gifted. Again, educational tests may measure abilities to which stupid pupils may be trained by dint of years of effort. Intelligence tests are supposed to measure transferred ability only. Again, bright pupils are not, as a rule, permitted to go forward as fast they could, consequently they

get no opportunity to learn the abilities measured by the educational tests because these abilities are only taught in higher grades. Finally, it is possible that the variability of E.Q.'s is less than the variability of I.Q.'s, since the variability of I.Q.'s has been increasing since birth, while the variability of E.Q.'s has only been markedly increasing since school entrance. If this last is true all low I.Q. pupils and all high I.Q. pupils would automatically have higher and lower E.Q.'s respectively, while average I.Q. pupils would have equal E.Q.'s.

Are Pupils Now Classified by Educational Age?—Teachers, educational tests, and mental age are in substantial agreement that pupils are not classified in homogeneous groups. The median educational age for each grade and section in School Y was as follows:

IIIA	IIIB	IVA	IVB	VA	VB	VIA	VIB	VIIA	VIIB	VIIIA	VIIIB
100	107	112	122	128	133	143	132	144	144	149	157

In School Y, VIB is actually behind VIA and there is practically no progress at all between VIA and VIIB.

But the position of the grade medians, improperly spaced as they are, does not begin to suggest how bad the classification really is. Fig. 2 permits a comparison of the amount of total grade overlapping. At a glance this diagram tells us that the extreme range of each grade is about 50 months in terms of educational age, which is equivalent to a range of about four typical grades, while the interval between the two grades is 2.5 months. The range of ability within one grade of School Y is then approximately 20 times the difference between two adjoining grades.

But before many assertions can be made about the amount of overlapping it is necessary to enquire which of the three possible causes of overlapping is responsible. These three causes are, (*a*) the trivial or inadequate nature of the mental traits measured, (*b*) the unreliability of the test, and (*c*) incorrect classification. Is the large amount of overlapping between grades of School X and School Y due to the last cause or to the first two causes?

The tests used in both schools were not of a trivial nature. They measured mental traits which are now and ought to be central in classification. If two groups ideally classified and separated for educational purposes were measured for weight, a large overlapping between groups would be found. But this would not be an indictment of the classification for variations in weight have little significance for educational

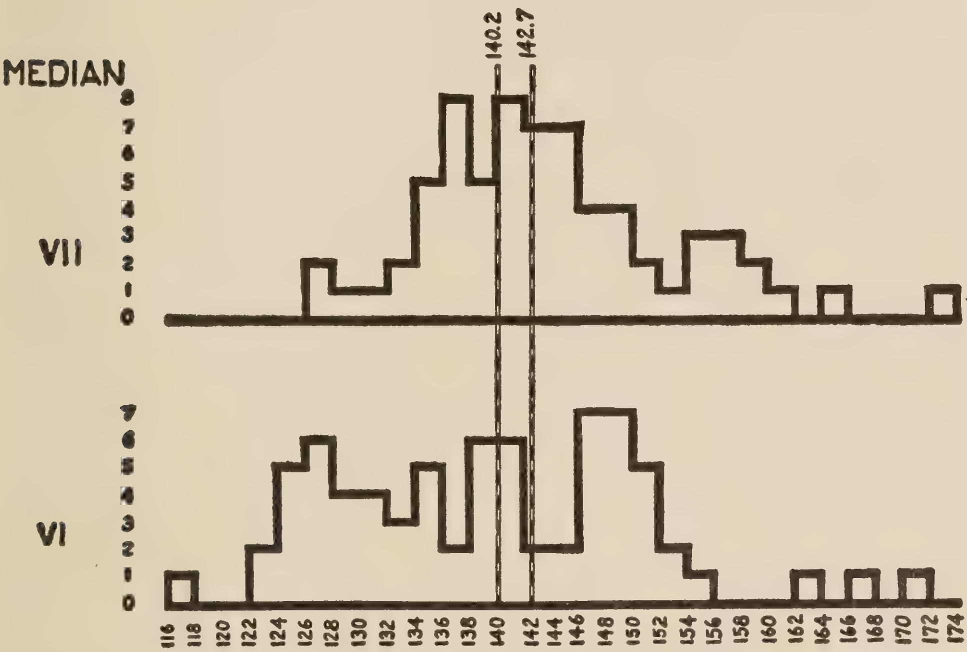


FIG. 2. A Graphic Picture of the Amount of Overlapping of the Educational Ages of Pupils in Grades VI and VII of School Y.

classification. The traits measured in these schools were, unlike weight, of vital significance for educational classification.

While the traits measured were not trivial they were probably inadequate. Kruse [10] has shown that the obtained overlapping of abilities from grade to grade is always greater than the true overlapping and hence any classification based upon inadequate measures is sure to be too drastic. If classification ought to be based upon methods-of-work abilities, a charge of inadequacy cannot easily be

[10] Paul Kruse, *The Overlapping of Attainments in Certain Grades;* Teachers College, Columbia University, New York, 1918.

placed against the tests of School X. The tests used are fairly representative samplings of the methods-of-work abilities. If classification ought to be based upon informational abilities such as are given by history and the like, the charge of inadequacy has some validity. It is easy, however, for those engaged in classification to meet this charge, for objective informational tests exist and these tests may be used if subsequent research reveals that there is not a high correlation between method abilities and informational abilities.

It is possible to go beyond Kruse and register a charge of inadequacy against all classification both old and new. Practically all classification, whether based upon objective tests or teachers' judgments, has considered abilities and has almost completely ignored purposes. And yet so long as purposes together with abilities make up the curriculum it may be reasonable to ask that pupils be graded upon their possession of purposes as well as their possession of abilities. Half, if not more, of what goes to the making of an educated individual is now practically ignored in every system of classification. Since any attempt at present to measure purposes is likely to be highly inaccurate perhaps it is well that they have been ignored. It is important, however, that they not be forgotten, and that scientific students of education continue their efforts to make classification measurements more adequate than is now possible.

Overlapping between grades is exaggerated not only through inadequacies but also through inaccuracies or unreliability of the measurements. Imagine two regiments, each of which is composed of men of identical heights and each of which differs from the other by a very small amount. If measurers who were ignorant of the present accurate classification were given crude instruments for measuring height and told to determine the height of the men, the results would show an overlapping between regiments due to mistakes of measurements. This factor is always operating in educational testing to make classification appear

worse than it really is. Kelley [11] gives the following formula for determining how much too large is the obtained variability and hence how much too large is the obtained overlapping.

True *S.D.* = Obtained *S.D.* $\sqrt{\text{Self-correlation coefficient.}}$

Thus were the self-correlation coefficient (later chapter) of all scores on all tests combined which were given to School X .25, i. e., were the self-correlation .25 between the composites or educational ages for all pupils in one grade and another similar determination of composites or educational ages, the true S.D. would be computed thus:

True *S.D.* = Obtained *S.D.* $\sqrt{.25}$ = .5 Obtained *S.D.*

That is, were the self-correlation no more than .25 the true overlapping is only .5 or one-half the obtained overlapping—the interval between grades is twice the obtained interval. The self-correlation for the tests used at School X is probably sufficiently high to make the obtained overlapping so nearly the same as the true overlapping as to require no correction. The self-correlation for the reading test alone is over .8. When all the tests are combined the self-correlation is certainly over .9, though it has not been computed, and is probably about .97. Substituting it is found that the obtained *S.D.* is almost the same as the true *S.D.*

True *S.D.* = Obtained *S.D.* $\sqrt{.97}$ = .98 Obtained *S.D.*

Classification and Placement Table for a Small School.—The small illustrative School X has been reclassified. The final disposition of each pupil is shown in the last column of Table 2. The first step in regrading this school was to space off the grade intervals. Only one of the many ways for doing this need be described. The median educational age for Grade III is 111 months. The median educational age for Grade VIII is 178 months. This gives (178 — 111) or 67 months to be divided into five equal portions for the five grade intervals. This gives 67 ÷ 5 or 13.4. It

[11] Truman L. Kelley, "The Measurement of Overlapping"; *The Journal of Educational Psychology*, November, 1919.

is reasonable to ask each of the grades from III to VIII to do its part in lifting the educational age of its children from 111 to 178 months. Each grade's part is 13.4 months.

But some may reasonably contend that the school should, at least, lift its pupils from their present third-grade position of 111 months educational age to the eighth grade norm of 180 months. That is, they should lift them $(180 - 111) \div 5$ or 13.8 months per grade instead of 13.4 months. This latter contention has not been accepted partly because to make the grade intervals 13.8 would necessitate a rather drastic reclassification of the pupils, and partly because it is known that this school could not well accomplish the task set for it without retarding its stupid pupils even more than at present. In sum, the past achievement of the school has been accepted as the basis of classification. Even their achievement of the past, like the achievement of the norm school, has been secured through holding back the stupid pupils while the brighter ones went ahead.

Since the classification of the pupils in School X should not entirely neglect E.Q., the second step has been to determine the median E.Q. of all the pupils. The median of the six median E.Q.'s for the six grades was sufficiently accurate for the purpose.

	III	IV	V	VI	VII	VIII	Median
Median E.Q......	90	97	105	102	97	103	100-

Knowing that the third-grade median educational age is 111, that the grade intervals are 13.4 and that the typical E.Q. of the school is 100, it was possible to construct a reclassification table for present pupils and a placement table for future pupils. In Table 4 each new grade median has been made larger than the preceding one by 13.4. To facilitate classification the quarter point below and above each grade median has been given. Since the tests were given at the extreme end of the school year, all pupils listed in Grade III, IV, etc., of Table 2 were just on the point of becoming Grade IV, V, etc., pupils respectively. And since it was desired to state, not in what grade each pupil

should be at the time the tests were given, but in what grade the pupils should be placed the following September, Table 4 was constructed as though the median educational age of 111 months belonged to the fourth instead of the third grade. When tests are given in the middle of a grade it is preferable to show, instead, the grade where the pupils should be at the time the tests are given.

TABLE 4

Reclassification Table for Present Pupils and Placement Table for Future Pupils of School X

	Ed. Age	E.Q.	Grade
Median	97.6 (est)	100	III
Third Quarter	101.1		
First Quarter	107.5		
Median	111.0	100	IV
Third Quarter	114.5		
First Quarter	120.9		
Median	124.4	100	V
Third Quarter	127.9		
First Quarter	134.3		
Median	137.8	100	VI
Third Quarter	141.3		
First Quarter	147.7		
Median	151.2	100	VII
Third Quarter	154.7		
First Quarter	161.1		
Median	164.6	100	VIII
Third Quarter	168.1		
First Quarter	174.5		
Median	178.0	100	IX
Third Quarter	181.5		
First Quarter	187.9		
Median	191.4 (est)	100	X
Third Quarter	194.9		
First Quarter	201.3		
Median	204.8 (est)	100	XI

Rules for Reclassifying Small School.—In reclassifying the pupils of School X, the following rules were adopted. These rules are not entirely arbitrary. They are based upon some experience with this method of reclassification.

1. *No pupil will be demoted or denied his normal promotion unless his educational age falls below the third quarter of the grade in which it is proposed to place him.* This rule gives the slow pupil the benefit of any doubt concerning the reliability of the tests, and is a slight concession to the idea that chronological age has some significance for classification. Further, it may not be wise to run the risk of discouraging by demotion or failure of promotion a stupid pupil who may be doing his best unless it is pretty certain that he really needs the work below.

2. *No pupil will be skipped over a grade whose educational age does not exceed the first quarter of the grade to which it is proposed to skip him.* If the classifier cannot afford to risk a few failures as a result of his recommendations he had better require that the pupil exceed the median of the grade to which he is promoted.

3. *No pupil whose E.Q. is below the typical E.Q. of the school will be permitted to skip one or more grades unless his educational age exceeds the median of the grade to which he is skipped.* If his educational age exceeds the first quarter of the grade to which he is skipped and his E.Q. is above the typical E.Q. of the school his relative position will very probably be higher still by the end of the year. If, however, his E.Q. is low he may be at the bottom of the class by the end of the year and thus be in danger of failing.

There are other considerations which ideally should influence classification. Besides an attempt to diagnose differences between E.Q. and I.Q. in order to discover which is the better index of rate of progress, it might be profitable to inspect each pupil's record on each test separately. A pupil might possibly have a high composite on all tests together with a sufficient disability in one to prove fatal. Arithmetic and reading should be carefully inspected. For

example, there is scarcely a pupil in Grade III of School X (Table 2) who does not have a zero score in subtraction. Why this is so is something of a mystery. This single disability might prove fatal to the arithmetic work of those pupils who have been skipped to Grade V. The double promotion was recommended only on condition that pupils having marked special disabilities be given special help.

The Reclassification.—The pupils of Table 2 have been reclassified according to the three simple, objective, and easily applied rules stated above, and in accordance with Table 4. The reclassification recommended is shown in the last column of Table 2. The first pupil in Grade III, namely pupil Ant, has been permitted to skip Grade IV and enter Grade V. His educational age of 122 months is above the first quarter of Grade V and his E.Q. of 112 is above the typical E.Q. of the school, namely 100. The next pupil, Bau, has been given his regular promotion to Grade IV. He exceeds its first quarter but his E.Q. is below 100. This would disqualify him from being skipped from Grade II to Grade IV, but does not rob him of his regular promotion. Pupil Soh has been kept in Grade III. His educational age is below the third quarter of Grade III and his E.Q. is as low as 71. Educationally he would probably be lost in Grade IV, and what is more important he would rob the regular pupils of Grade IV by demanding an undue amount of the teacher's time, etc. Pupil Mye of Grade IV has been skipped over Grades V and VI. His educational age is even beyond the third quarter and his E.Q. is 114. The child has been given, not a favor, but simple justice. Pupil Sim in Grade V is a doubtful case. He did not get the benefit of the doubt. He is still young and the school as a whole is slightly below standard. Pupil Fra in Grade VI shares with Pupil Kim of Grade III the honor of being the most intelligent child in the school. He is probably a genius. His educational age and E.Q. qualifies him for Grade X. Although the table shows him promoted to Grade X it was not actually recommended that any child be promoted be-

yond the first year of high school, partly because high school pupils are of a better intellectual caliber than elementary school pupils, partly because high school work is not a continuation of elementary school subjects, and finally because the recommendation would in all probability have been utterly disregarded by the high school to which the pupil would go. Pupil Kim in Grade VI is also advanced to high school. He is a near genius and is a brother of pupil Kim in Grade III. Pupil Ant in Grade VI has been skipped over Grade VII just as his brother pupil Ant in Grade III was skipped over Grade IV. Among others, pupils Bel, Lut, and Hen in Grade VII have been sent to high school. They have completed the elementary school work better than the typical pupil. The action taken in these cases may well be questioned, because their E.Q.'s are below 100. They will meet a much fiercer competition in high school. Another year in the elementary school will probably not, however, improve their high-school chances. Since they are over fifteen years old there is little hope for much further growth in mental age. Pupil Pug in Grade VII has close to the highest educational age in the school. He has been skipped to high school just as his sister, pupil Pug, in Grade VI was skipped over Grade VII. Just as two watches, built at different times by the same master-hand, run together, so these brothers and sisters, springing from a common heredity, tend to move at closely similar rates through the school.

In case it is desired to reclassify the school on the basis of what typical schools are achieving rather than upon the basis of the past achievement of the particular school in question the classification table should be derived, not from the median achievement of the grades in the particular school, but from Table 3. All pupils whose educational ages fall below $\frac{89 + 102}{2}$ are Grade I pupils and should be placed in Grade II, those between $\frac{89 + 102}{2}$ and $\frac{102 + 115}{2}$ are

Grade II pupils and should receive their promotion to Grade III, those between $\frac{102 + 115}{2}$ and $\frac{115 + 128}{2}$ should be placed in Grade IV, those between $\frac{115 + 128}{2}$ and $\frac{128 + 141}{2}$ should be placed in Grade V, and so on. Decision in doubtful cases should depend upon the size of the pupil's E.Q. In case half-yearly promotions are planned, pupils should be placed in the lower half of Grade V whose educational ages fall between $\frac{115 + 128}{2}$ and 128, and in the upper half of grade whose educational ages fall between 128 and $\frac{128 + 141}{2}$ and similarly for other grades.

The Amount of Promotion and Demotion Necessary.—The amount of reclassification necessary in this sample school, even when the classification has been somewhat conservative, is shown in Table 5.

TABLE 5

Distribution of Changes Made in Reclassifying School X by Means of Educational Tests. (Data from last column of Table 2)

Amount of Change	*Number of Pupils*						
	III	IV	V	VI	VII	VIII	Total
Demoted three grades	0	0	0	0	0	0	0
Demoted two grades	0	0	0	2	2	0	4
Demoted one grade	2	1	1	1	3	3	11
No change	13	13	11	8	7	5	57
Promoted one grade	4	1	2	5	5	1	18
Promoted two grades	0	2	0	1	1	2	6
Promoted three grades	0	0	0	1	1	0	2

Table 6 gives similar data for a larger school—School Y—where the technique of reclassification was practically the same. The tests used in this school were Thorndike's Reading Scale Alpha 2, Thorndike's Visual Vocabularly Scale

A_2 x, Trabue's Completion Scales B and C, Ayres' Spelling Scale, Starch's Reasoning Test in Arithmetic, Woody-McCall's Mixed Fundamentals, Parts I and II, and Composition.

TABLE 6

Distribution of Changes Made in Reclassifying School Y by Means of Educational Tests

Amount of Change	*Number of Pupils*						
	III	IV	V	VI	VII	VIII	Total
Demoted four grades	0	0	0	0	0	0	0
Demoted three grades	0	0	0	0	0	0	0
Demoted two grades	0	0	0	0	0	1	1
Demoted one grade	0	4	1	2	5	13	25
No change	39	39	38	29	45	36	226
Promoted one grade	15	16	39	27	12	9	118
Promoted two grades	0	5	9	11	3	1	29
Promoted three grades	0	0	0	1	1	3	5
Promoted four grades	0	0	0	2	0	1	3

The two tables when combined lead to the following conclusions which need to be only slightly discounted because of unreliability of the tests:

1. About 44 per cent of pupils are wrongly classified.
2. About 34 per cent of pupils are misplaced one grade.
3. About 10 per cent of pupils are misplaced two or more grades.
4. Only about 8 per cent of pupils are pushed ahead of the grade where they belong, while nearly 36 per cent are held back from the grade where they belong.

Success of the Reclassification in School X.—The reclassification of the pupils in School X as shown in Table 2, was recommended on the following conditions: (1) All promotions and demotions were to be trial promotions and demotions, and the pupils were to be so informed. (2) After four weeks of trial, the principal in consultation with the teachers was to construct a series of examinations upon the material studied during the four weeks and try these

tests upon the pupils. (3) The teachers were to rank the pupils in their respective classes upon the quality of their work during the four weeks. (4) The principal and teachers were to decide the final disposition of the pupils.

The demoted pupils have "made good," i. e., there has been no disposition to question the recommendations in their cases. Not one has been returned to his original grade. What happened to the promoted pupils for whom reports are available is shown in Table 7. This table is read as follows: Pupil War J., who has an E.Q. of 90, was promoted over Grade IV. He ranked, according to the educational tests, first among the sixteen pupils who, together with him, made up Grade V. He ranked first among the same sixteen

TABLE 7

What Happened to the Specially Promoted Pupils of School X

Pupil	E.Q.	Grade Skipped	Rank by Ed. Age	Rank by Principal	Rank by Teacher	Final Disposition
War J.	90	IV	1-16	1-16	6-16	V
War R.	108	IV	4-16	2-16	3-16	V
Kim	134	IV	10-16	7-16	9-16	IV
Ant	112	IV	12-16	8-16	11-16	IV
Mye	114	V & VI	1-16	3-16	5-16	VII
Sco	125	VI	3-16	11-16	14-16	VI
Van	110	V & VI	7-16	.....	13-16	VI
Hoy	112	VI	10-16	7-16	9-16	VII
Fra *	135	VII	1-16	4-16	6-16	VIII
Kim *	131	VII	4-16	10-16	9-16	VIII
Spi	118	VII	7-16	14-16	13-16	VII
Lan	108	VII	9-16	16-16	16-16	VII
Pug	110	VII	10-16	12-16	14-16	VII
Ant	121	VII	11-16	.....	11-16	VII
Mit	113	VII	12-16	8-16	15-16	VII
Pug	126	VIII	3-10	1-10	5-10	IX
Average			6.2-15.6	7.4-15.6	9.6-15.6	

* Certain difficulties kept these pupils from being sent to high school as recommended.

pupils who were tested by the principal upon four weeks of school work. In the judgment of his teacher he ranked sixth. He was finally retained in Grade V.

Table 7 permits an interesting psychological study of the pedagogical mind. The data are too inadequate to allow other than tentative generalizations, in order to provoke further study. Anyway, the purpose here is not so much to draw conclusions as to describe a process. The table suggests the following:

1. A specially promoted pupil tends to be ranked lower by the teacher's judgment than by the principal's examination or by standard educational tests. The averages at the bottom of the table show that the average ranks by tests, principal, and teacher are respectively 6.2, 7.4, and 9.6 out of about sixteen pupils.

2. A young, specially-promoted pupil must succeed beyond a shadow of doubt or he will be demoted. Pupils Kim and Ant of Grade V, and possibly Mit of Grade VIII, did better than was originally anticipated and yet they were reduced a grade.

3. A pupil's educational age and E.Q. must at least exceed the median of the grade to which he is sent or the teacher and principal will probably return him. And it should be remembered that the principal and teachers of School X were friendly to the experiment.

4. The school's staff is convicted of injustice by its own measurements. Can anyone unacquainted with school traditions give a rational explanation of why pupils Kim and Ant were sent back to Grade IV? The real fact is that these teachers require a young pupil to do, not the typical work of the grade, but the *best work* in the grade. The teachers of School X testify that most of the young pupils demoted had rapidly risen in rank since the opening of school. With their high E.Q.'s it is probable that this process would continue throughout the year, thus making their class status better and better.

The teachers explained that pupils Kim and Ant were

demoted, even when their rank was satisfactory, because those ranking below them were relativly stupid pupils. This factor would not have influenced the teachers had anyone been present to explain that while these pupils were stupid, they were also much over age. Additional years of schooling had balanced their stupidity. While their E.Q.'s were low their educational ages were as high as pupils Kim and Ant. If the measurements of the principal and the judgments of the teachers be accepted at their face value, only one pupil was legitimately sent back and that is pupil Lan. All the others have paid the penalty of their prominence and particularly of their unfortunate youthfulness, unless it be assumed that the fundamental basis for the classification of pupils should be chronological age.

What happened to the pupils who were sent back? If the effect of demotion is to produce sulkers, special promotion should not be given unless there is considerable certainty that the promotion will be maintained. The principal reports that one or two were glad to go back to their former companions, some did not want to go back, some didn't care, every pupil except Lan are at the top or near the top of the grades to which they were returned. The principal reports that those originally demoted on the basis of the tests are happy and satisfied.

Franzen has since tried the experiment in the Garden City elementary school of giving special promotion only to those pupils whose educational age and E.Q. or mental age and I.Q. both exceed the median of the grade to which they are sent. In no case was a pupil afterward demoted.

Procedure for Reclassifying Large School.—The procedure recommended for reclassifying a small school has one fundamental defect. It secures homogeneity of educational status, but it does not guarantee equal rate of progress. E.Q., the prophet of future progress, was used only incidentally.

A thoroughgoing use of both educational age (or mental age) and E.Q. (or I.Q.) requires a school with enough pupils

and teachers to make two or, preferably, three or more classes per grade. Given this condition it is possible to have parallel groups, some of which either progress more rapidly through the grades or else take a wider educational swath. This enables the school to keep together pupils with like educational ages, E.Q.'s, chronological ages, and social and vocational needs. It insures that no pupil will be forced to skip vital parts of the school's work in order to progress rapidly, that no pupil will lack the training which comes from keen competition, that no pupil's vanity will be fostered by a perception of unquestioned superiority on his part, that no pupil will form habits of mental laziness by living in a mentally non-stimulating atmosphere and by being continually called upon to master difficulties which he has already mastered, and, finally, that no pupil will be persecuted or taught social timidity by the brute physical strength of the chronologically older. Many of these advantages can be secured through a classification on the basis of educational age alone. All can be secured through a classification by both educational age and E.Q.

The steps in the procedure of classifying both by educational age and E.Q. follow:

1. Pupils should be classified into the various grades on the basis of educational age without regard to E.Q. It is suggested that no pupil be specially promoted or demoted unless he exceeds or falls below respectively the first quarter or third quarter of the grade in question, as was done in the case of School X.

2. When the horizontal grade classification on the basis of educational age has been completed, the pupils in each grade should be divided into groups on the basis of E.Q. If, for example, there have been enough pupils placed in the third grade by the first classification to make three classes, all pupils whose E.Q. is, say, over 110 may be placed in one group, all whose E.Q. is between 90 and 110 in another group, and all whose E.Q. is below 90 in the third group. This process should be repeated for each grade.

The E.Q. points selected will depend upon the distribution of the E.Q.'s in each grade, and the number of pupils desired in each group. The usual practice is to place fewer pupils in the superior and inferior groups than in the normal groups. Or instead, the pupils in each grade may be arranged in order according to the size of their E.Q.'s. The highest third can be placed in one group, the middle third in another, and the bottom third in another.

The procedure described for both a small school and a large school is identical whether educational or intelligence tests are being used. If intelligence tests are being used, mental age takes the place of educational age, and I.Q. the place of E.Q. If pupils are being classified into grades only and not into sections within each grade, neither educational age nor E.Q. are absolutely required. Pupils could be grouped with reasonable accuracy on the basis of their composite scores. In this case composite score takes the place of educational age in the classification table.

The Placement of Future Pupils.—Copies of the tests used in classifying the pupils of School X were left with the principal, together with standard instructions for applying and scoring them. When a new pupil arrived at the school, the principal applied these tests, scored them, multiplied the score on each test by the same multipliers shown at the bottom of Table 2, added the weighted scores to get the pupil's composite score, converted his composite score into educational age according to Table 3, divided the educational age by the pupil's chronological age to get E.Q., placed the pupil in the appropriate grade according to his educational age, his E.Q., and Table 4, after having made due allowance for the progress of the regular pupils since they were reclassified.

III. Classification by Teacher's Judgment

Inadequacy of Teacher's Judgment.—The teacher's judgment is an inadequate basis for classification (*a*) when

pupils first come to school, whether from the home or from some other school, and (*b*) even when pupils first enter the class from some other class in the same school. It is difficult for a teacher to give a satisfactory judgment even after considerable experience. Teachers lie awake nights worrying whether or not to promote a given pupil partly because they are somewhat uncertain of their judgment. No judgment at all is available for pupils with whom the teacher is unacquainted.

Again, the teacher's judgment is inadequate even when she knows her own pupils well. Teachers are not blind. They can distinguish between their brightest and dullest pupils with considerable accuracy. But lines of classification are rarely drawn through a single class. These lines usually break across several classes. A teacher may know her own pupils well and yet be unable to tell whether her ablest pupils are equal to, superior to, or inferior to the average or brightest pupil in some other class. Some form of measurement is needed which does not require a previous acquaintance with pupils and which compares pupils in different classes with as great ease as it compares pupils in the same class.

Inaccuracy of Teacher's Judgment.—The very nature of subjective measurement and of the pupils being measured make a teacher peculiarly liable to err in estimating the ability of pupils. Teachers' judgments are subject to certain constant errors. One such error is the tendency to confuse conduct with achievement. An estimate of a pupil's ability in reading is usually all tangled up with a cheery "good morning," a courteously opened door, close attention to the teacher's remarks, and other examples of impeccable behavior. It is probable that conduct (purposes) should be considered in classification, but conduct should not be allowed to cloud a teacher's estimate of abilities.

Again, Terman finds that teachers frequently err in estimating intelligence because they fail to take age and emotional differences into account. An over-age or emotionally

vivacious pupil is usually rated too high and an accelerated or phlegmatic child is usually rated too low. Finally, Whipple finds after an unusually exhaustive study of gifted children, that while teachers are not likely to rate dull pupils as average or superior, they are likely to rate superior children as average. He therefore concludes that mental examiners are as much needed for selecting superior children as for selecting inferior children, because these mental examiners employ measuring instruments which are not subject to these constant errors.

Importance of Teacher's Judgment.— Teachers' marks are important because they are now and will continue for some time to be the most universal method of rating pupils. In fact, they may continue forever to be the criterion for classification, because teachers will soon be familiar with the simple mysteries of scientific measurement. They will themselves use tests with the same ease and fluency that they now use text-books. More and more they will base their judgments upon objective rather than subjective measurement. When this time arrives teachers' marks will be not only as accurate as objective measurement, but they will be objective measurement plus something else.

This something else makes teachers' marks valuable even apart from objective tests. There are significant segments of each pupil's make-up which tests do not now touch. In occasional instances this segment has more importance for classification than the segment measured by tests. Teachers' judgments appear to be the only immediate hope of measuring pupils' purposes. Furthermore, the teacher can weight physical, emotional, and social characteristics in ways that tests cannot. In so far as these elements in the make-up of a pupil are aims of instruction, and in so far as they *are improvable by school instruction,* they ought to be weighted in determining promotions. Hence the teacher's judgment should be a factor in determining promotion, at least until the time arrives where these relatively intangible abilities are shown not to exist, or to be included

in the score on objective tests, or shown to be abilities which are unimprovable by school instruction.

Computation of Pedagogical Age.— The intelligence test determines mental age, the educational tests determine educational age, the status of a pupil as judged by a teacher may, for convenience, be called pedagogical age. But how can a teacher whose rating of pupils is confined to her own class and is relative to her own class determine a pedagogical age which is comparable to mental and educational age and may be combined with them to decide a pupil's classification?

This is a problem which has been worrying psychologists and statisticians for many years. Many schemes have been proposed to increase the fairness and usefulness of teachers' marks. To be thoroughly satisfactory whatever plan is proposed must provide that a teacher's marks for pupils in one grade will be comparable with another teacher's marks for pupils in another grade or another class in the same grade. Allowance must be made for the fact that some classes are large and some are small, that some have unusually stupid pupils while others have unusually bright pupils, and that some classes have a large variability while others have a small variability. Finally the plan must yield scores comparable with scores on tests, and which enable the school to establish permanent limits of ability for each grade and class for purposes of classification and placement.

I have devised a simple scheme of marking which will meet fairly well all the above conditions. This plan follows:

1. Reliably measure with some objective test or tests all the pupils of the school. This may be an intelligence test, a battery of educational tests, or a single educational test. If only one educational test is used perhaps the best would be one which measured the pupils' comprehension in silent reading.

2. Compute mental age or educational age (or reading age) for each pupil.

3. Arrange the pupils in each class separately in order of their mental age or educational age on the objective tests.

4. Have each teacher rank the pupils in her class in order of how they should, in her judgment, be promoted. If several teachers rank the same pupils their ranks may be averaged. By multiplying one teacher's ranks by two or three that series of ranks may, if desired, be given double or triple weight.

5. Give the pupil ranked highest by the teacher the highest mental or educational age made by any pupil in her class on the objective test. Give the pupil ranked second by the teacher the second highest mental or educational age made on the objective test. These are the pupil's pedagogical ages. Thus the test bridges, for the teacher, the gap between grades and classes. The test shows the teacher where to begin rating, and how much variability to provide for.

While a test will make occasional mistakes in the case of individuals, their determination of differences between groups and the variability of a group is approximately correct, particularly if enough tests are given to make results adequate and reliable. It is probable that the true variability of a group and the true differences between groups in mental age or educational age are about the same as the true variability and true differences, respectively, in the traits upon which the teacher bases her judgments. There may be a small error at the extremes of each class. Thus tests help the teacher where she most needs help, namely, in relating her pupils to pupils in other classes and grades. Tests tell her what scores to assign to the pupils in her class, but leave her free to decide who shall receive these scores. Pupils whose educational age and pedagogical age differ widely should receive special attention from the school psychologist.

6. Average the mental or educational age and the pedagogical age to get each pupil's promotion age. Or the

pedagogical age alone may be used for a promotion age. If each pupil has a mental, educational, and pedagogical age, then all three may be averaged, giving any desired weight to each. The weight given will depend upon the number and nature of the tests used.

7. Compute the Promotion Quotient for each pupil. The top pupils in a sample class are given below:

Pupil	Ch. Age	Ed. Age	Rank by Teacher	Ped. Age	Promotion Age	Promotion Quotient
r	140	144	2	138	141	100.7
a	150	138	3	133	135.5	90.3
h	110	133	1	144	138.5	125.9
b	121	130	4	130	130	107.4
etc.	etc.	etc.	etc.	etc.	etc.	etc.

8. Classify pupils into grades on the basis of promotion age and into sections on the basis of Promotion Quotient.

9. Since the Promotion Quotient shows the pupil's typical rate of progress a little simple arithmetic will show what the promotion age of each pupil may be expected to be after a year of instruction. This expected age may, if desired, take the place of tests, but not of teacher's ranking, at the next promotion. This avoids the use of tests each year except for new pupils.

Teachers' judgments were not utilized in the reclassification of School X because the above plan is a subsequent discovery.

Objections to Classification by Tests.—By way of summary it may be well to list the more common objections to promotions on the basis of educational or intelligence tests.

1. *Young pupils are forced to compete with the mentally more mature.* This is a relic of the old notion that all pupils are born equal and that subsequent mental age keeps pace with chronological age. In general this objection represents a misplaced sympathy. Every investigation shows that it is a rule for the young pupils to be leading their classes and for the older pupils to be struggling to keep up. Classifica-

tion by both educational age and E.Q. entirely meets this objection.

2. *Young pupils have difficulty in making social adjustments.* It would be truer to say that older pupils have difficulty in adjusting to the younger ones. There is an undoubted tendency for older pupils to dislike the presence of a much younger pupil, for his presence is a standing insult to their intelligence. How serious this jealousy is needs to be investigated. Classification by both educational age and E.Q. completely meets this objection.

But what will the gifted pupils do when they reach the high school while still very young? One suggestion is that they delay their arrival at the high school by taking a wider educational swath. If, however, the curriculum has been properly constructed this means that the gifted pupil will be spending almost half his time upon material of relatively small value. The only satisfactory solution is to provide a path for the geniuses which leads from the first grade through the university, so that the genius pupil may be with his kind throughout his entire educational career. If he graduates from the university while still too young he may be employed in national research or on other large social enterprises until he is judged sufficiently mature physically to take his place in the general social group.

The only visible solution for the small school is to promote the young gifted pupil just as often as the older pupils will permit without making his life miserable. Another solution is to abolish the school which is too small to make adequate provision for individual differences among its pupils and to substitute the consolidated school in its place.

3. *Causes vital gaps in pupil's education.* Classification by educational age meets this objection provided the testing has been thorough. To receive a high educational score shows that these gaps have somehow been filled. If the educational testing has been inadequate pedagogical age may be included.

It is difficult to believe that this is the real objection. It

can be demonstrated that older pupils have phenomenally large gaps in prerequisite abilities. But this does not seem to produce any particular concern. Worry comes only when the young pupil is involved. Educators find it almost as difficult as laymen to prevent themselves from thinking in terms of such irrelevant surface factors as chronological age, physical size, and brute muscles. We think of children as we would of elephants or dinosaurs. Considering how much of our lives are regulated by chronological age, this is not surprising. We are born at zero years of age, compelled to begin school at six, permitted to leave school at fourteen, allowed to marry at sixteen, entitled to vote at twenty-one, and are given an average salary at that age where a long life of usefulness is passing into decline. Squeezed in between a chronological end and a chronological beginning the passage through school naturally becomes a chronological procession.

4. *Disregards health.* Some picture the gifted child as a frail, forced, hot-house flower. Terman, after a careful study of many gifted pupils, concluded that they were no more frail than ordinary children. He found that some were frail and some robust. Consequently if there is any reason to suppose that health will be sufficiently improved by giving the pupil intellectually easy tasks, health should certainly be considered.

There is, however, a fear abroad that a pupil's mind may, like Jefferson's Constitution, be stretched until it cracks. In his Columbia University Master's thesis Franzen describes a ten-year-old pupil in Grade V with an I.Q. of 178. This genius distinguished between poverty and misery, thus: "Poverty is the lack of things we need, misery is the lack of things we want." He defined a nerve as the "conduction unit of sensation" and explained correctly what he meant thereby. It was discovered that he had read all the textbooks of the grades ahead of him. His two able parents after a consultation with the family physician refused permission for him to be promoted from the grade where he

was bored almost to extinction because it might *strain his mind!*

5. *Emphasizes the intellectual to the exclusion of character traits.* This is another way of saying that pupils are classified by their abilities and not by their purposes. It may be that a pupil can be taught desirable purposes in one grade as easily as in another. It is certain that purposes do not fall into such close hierarchies as do abilities. Furthermore, it is possible that most pupils who are promoted for their intellectual achievements would likewise be promoted for their composite character status. Terman[12] studied the extent to which intellectually gifted pupils possessed the following intellectual and personal traits: sense of humor, power to give sustained attention, persistence, initiative, accuracy, will power, conscientiousness, social adaptability, leadership, personal appearance, cheerfulness, coöperation, physical self-control, industry, courage, dependability, self-expression through speech, intellectual modesty, obedience, popularity among fellows, evenness of temper, emotional self-control, unselfishness, and speed. Any reader would not complain of any lack if he possessed intelligence plus this galaxy of traits. Terman found that all these traits correlated positively with intelligence, that is to say, with ability primarily. The first trait, sense of humor, has, in the case of gifted children, a correlation of .58. The last trait, speed, correlates .28. The others gradually vary between these extremes in the order named. Terman claims that he can roughly predict I.Q. from an average of these 24 traits.

The most common explanation given by teachers for the failure of certain specially promoted pupils to do satisfactory work is that they do not try. It is generally admitted that they could do the work of the grade if they only would. These remarks by teachers suggest two questions. (1) Since tests reveal that these pupils have actually mastered, some-

[12] L. M. Terman, *The Intelligence of School Children*, p. 58; Houghton Mifflin Co., New York City, 1919.

where, somehow, large segments of the curriculum, is it not possible that they are mastering the material of the new grade with such unobtrusive ease as to deceive even the keenest observer? (2) If the teachers are correct may not the pupil's lack of industry be due to improper habits formed by previous improper classification where industry was not required?

Classification by both educational age and E.Q. or mental age and I.Q. will meet some of the above objections. The inclusion of pedagogical age will meet others, notably the last. Where the school is so small that it is impossible to classify by both educational age and E.Q. it is necessary to weigh such of the above objections as are relevant in the case of a particular pupil against the likelihood, if he is not promoted, that he will develop habits of inattention, intellectual laziness, and disorderly conduct. On all these questions more research is sorely needed. Until fuller knowledge is available it behooves all users of tests for purposes of classification to exercise great care.

CHAPTER III

MEASUREMENT IN DIAGNOSIS

I. Diagnosis of Initial Situation

Plan of Discussion.—A principal or supervisor who assumes responsibility for a new school, or a teacher who takes charge of a new class is faced with the necessity of making two types of diagnoses: a general diagnosis of the initial condition and a more detailed diagnosis of the particular defects of classes or pupils.

In order to give concreteness to the discussion in this and the two succeeding chapters I shall keep in mind the teaching of reading. A reasonably competent student will be able to transfer the techniques described to other subjects. Marginal numbering will be made continuous

TABLE 8

Shows the Information Needed to Guide Instruction in Reading

Pupil	A	B	C	D	E	Mean
1. Chronological Age	121	123	124	118	100	
2. Initial Reading Score	40	43	30	31	38	36.4
3. Initial Reading Age	121	130	93	96	116	
4. Initial Reading Quotient	100	106	75	81	116	
5. Initial Mental Age	121	150	83	100	111	
6. Intelligence Quotient	100	122	67	85	111	97
7. Initial Accomplishment Quotient	100	87	112	96	105	100
8. Estimated Final Reading Age	131	142	100	105	127	
9. Estimated Final Mental Age	131	162	90	109	122	
10. Reading Score Objective	43	47	33	34	42	39.8
11. Final Reading Score	43	50	34	36	43	41.2
12. Final Reading Age	130	150	104	110	130	
13. Final Accomplishment Quotient	99	93	116	101	107	103
14. Final A.Q. Minus Initial A.Q.	—	—	—	—	—	+3

through the three chapters in order to indicate the steps in the total process of using tests in teaching.

1. ***Determine Each Pupil's Chronological Age.***—Line 1 in Table 8 shows the chronological age in months of five selected fourth-grade pupils. (These five pupils will be treated hereafter as though they constituted an entire fourth-grade class.)

2. ***Determine the Initial Ability to Comprehend What Is Read Silently.***—The Thorndike-McCall Reading Scale may be used to determine each pupil's initial ability to understand what he reads. An easy and difficult portion of this scale, together with the directions which accompany it, are reproduced below.

Read this and then write the answers. Read it again if you need to.

Nell's mother went to the store on Water Street to buy ten pounds of sugar, a dozen eggs and a bag of salt. She paid a dollar in all. Nell and Joe went with her. On the way home on Pine Street, they saw a fire-engine with three horses.

1. Was the salt in a box or a bag or a can or a dish?..................
....
2. How many eggs did she buy?...............................
3. What did the children see on Pine Street?..........................
....
4. What street was the store on?...............................

Read this and then write the answers. Read it again if you need to.

COLERIDGE

I see thee pine like her in golden story
Who, when the web—so frail, so transitory,
The gates thrown open—saw the sunbeams play
With only a web 'tween her and summer's glory;
Who, when the web—so frail, so transitory,
It broke before her breath—had fallen away,
Saw other webs and others rise for aye,
Which kept her prisoned till her hair was hoary.
Those songs half-sung that yet were all divine—
That woke Romance, the queen, to reign afresh—
Had been but preludes from that lyre of thine,
Could thy rare spirit's wings have pierced the mesh
Spun by the wizard who compels the flesh,
But lets the poet see how heav'n can shine.

30. Who acted like a spider? ....

31. Who or what is compared with the woman? ....

....

32. Copy the first word of the line which implies there had not been a continuous stream of like songs? ....

....

33. Complete the following with one word only:

"Those songs" really means those ....

Directions for Using Thorndike-McCall Reading Scale

Form 2

HOW TO APPLY TEST

To the Examiner: Distribute test booklets, with front page up. Request pupils not to open booklets until the signal is given. Have pupils fill out the blanks at top of the front page. Read the paragraph aloud while pupils read silently. Read the first question aloud. Have it answered orally and then in writing by the pupils. Stop pupils thirty minutes after saying *Open paper! Begin!* Give no further help.

HOW TO SCORE TEST

It is suggested that the examiner answer the questions on the first test page, study the scoring key below, score the first test page for the entire class, and then repeat the process for the other test pages. The four questions on the front page should not be scored. The scoring key gives correct answers, in some instances followed by incorrect answers. The only portion of the answer required for correctness is in italics. The scoring key is by no means exhaustive. Other answers are to be scored as correct or incorrect in accordance with the following suggestions:

1. Any answer equal in merit to the worst answer called correct in the key is to be scored correct.
2. Complete sentences are not required.
3. Call wrong a correct answer plus an incorrect or irrelevant answer.

4. Call correct a correct answer plus a harmonious addition.
5. Call incorrect any misplaced, omitted, or illegible answers.
6. Do not call an answer wrong because of alteration, use of abbreviation, or errors in capitalization, punctuation, spelling, or grammar unless any of these indicate actual misreading.
7. In general the pupil's answer must show that he has correctly read both the paragraph and the question and is able to word a readable response to both.

SCORING KEY

The only portion of the answer required for correctness is in italics.
* Incorrect answers.

1. The salt was in a *bag.*
2. She bought a *dozen* eggs; 12.
3. The children seen a *fire-engine* with three horses.
4. *Water* Street; *Walter* Street.
5. He *killed* the *fox; Kill it.*
   * Killed; Shot him.
6. The rabbit was *brown.*
7. *In* the *woods; In* the *forest.*
   * Woods.
8. Beauty had *three* brothers; *Three* boys; 3.
9. *Yes; Younger* than both; She was *younger.*
10. *No; No, Beauty;* The *youngest.*
    * Beauty: Young.
11. *Jealousy; Jealous.*
12. *Music* lessons; *Piano lessons; Play piano.*
    * Grace's music; Singing; Piano; Music and hens.
13. *No;* Grace *sell* them; *Not.*
14. *Yes; Less; He doesn't like* it *at all.*
    * Fred does not like music.
15. *Richard; Richard* and Edward.
    * Edward; Richard and Henry.
16. *Sam and Henry; Sam* and *Edward.*
    * Edward doesn't like Sam; Sam, Henry, Edward.
17. *Arthur* and *Richard.*
18. *Henry* and *Edward; Henry; Edward.*
19. One *Hour;* From *ten minutes of seven to ten minutes of eight;* He *left* the *house at ten minutes of seven* and *reached* the *store at ten minutes of eight.*
    * Almost an hour.

20. *Once* a week; *One time; One day; Every Sunday; Only Sunday; On Sunday; Sundays; At Sunday; 1 each week; 1 a week.*
* Sunday; One.

21. He rose, *dressed, ate breakfast,* and left house.
* Dressed and left house; Got ready for work; Rose, dressed, ate breakfast, and worked.

22. *No.*

23. *Farmers too prosperous* to trouble about it; Because the *men are so rich.*
* Too rich to take the trouble; Because harvesters not careful; Farmers are getting prosperous.

24. Out *in* the *wheat fields* of Kansas; *In Pawnee County; On* the *fields where wheat was left.*
* In harvest field; Gathering wheat; Going over wheat fields; In a wheat field in Pawnee County.

25. *After* the *planting.*
* After planting is done and weeds are killed; When dry or weedy; After the crop begins to grow.

26. *Plowing planting sowing weeding watering loosening* (any two); *Plows* and *plants.*
* Plowing hoeing; Plowing spading; He digs and plants; Loosened and watered.

27. *Shyness* and tendency to *stammer; Stuttering* and *shyness;* Retiring *nature* and *speech; Retiring stammering;* He *stammered* and *shy.*
* His shy and retiring nature.

28. On *mathematics; Mathematics* and preaching; Lectured on *mathematics.*
* Lecturer on mathematics.

29. He has *written* funny *books* and poems and made lantern pictures; Studying, *writing books;* Narrative *books;* He *was a writer; Alice in Wonderland.*
* Wrote.

30. The *wizard; Disease; Illness.*
* The lizard; Spun by the wizard.

31. *Coleridge; Thee; Poet; Life of* the *poet; One to whom* the *poem is addressed.*
* A man was compared; The person whom the author is writing about; Coleridge's writings.

32. *That.*

33. *Poems; Verses; Lines; Stanzas.*
* Pieces of poetry; Lyrics.

34. *Methods operation; Method operation; Two methods.*
* Two methods, operation; An operation, methods; An operation; Experiment operation.

35. *Separation resolution.*
* Separation dissemble.

HOW TO COMPUTE PUPIL SCORE

To compute a pupil's score, first total the number of questions he answers correctly, look up this total in the first

column of Table 9, and read in the second column the corresponding T score. The T score is the pupil's score and should be tabulated as in Column III of Table 12.

TABLE 9

Questions Correct	T Score	Questions Correct	T Score	Questions Correct	T Score	Questions Correct	T Score
0	23	9	33	18	43	27	63
1	25	10	34	19	45	28	67
2	26	11	35	20	47	29	71
3	26.5	12	36	21	49	30	76
4	27	13	37	22	51	31	79
5	28	14	38	23	53	32	82
6	29	15	39	24	56	33	86
7	31	16	40	25	58	34	92
8	32	17	42	26	60	35	96

HOW TO COMPUTE CLASS SCORE

The class or grade score is the arithmetic mean of the pupils' scores, as shown beneath Column III in Table 12.

HOW TO COMPARE CLASS OR GRADE WITH GRADE NORM

Find the appropriate grade norm in Table 10, and write it beneath the class or grade score as shown in Table 12. The norm for IV B is shown in Table 12 because the tabulation is for a IV B class.

TABLE 10

GRADE NORMS 2A-7B

At the End of	2A	2B	3A	3B	4A	4B	5A	5B	6A	6B	7A	7B
Mean Norm	26	30	33.7	37.3	39.6	41.8	44.9	48.0	50.9	53.7	56.0	58.3
Approx. No of Pupils in Thousands	0.2	0.3	3	5	5	10	5	10	5	10	5	10

GRADE NORMS 8A-12B

At the End of	8A	8B	9A	9B	10A	10B	11A	11B	12A	12B	Superior Teachers
Mean Norm	59.6	60.9	61.5	62.1	62.9	63.6	64.5	65.4	66.8	68.1	72
Approx. No. of Pupils in Thousands	5	10	Est.	1	Est.	1	Est.	1	Est.	1	0.3

HOW TO COMPARE PUPIL WITH AGE NORM

To compare each pupil with age norms the examiner should look up the pupil's T score in the first column of Table 11, find in the second column the corresponding *Reading Age,* and divide the Reading Age so found by the pupil's chronological age to find his *Reading Quotient,* as shown in Columns I, IV, and V of Table 12. The Reading Quotient is 100 for the child with normal reading ability, and proportionately below or above 100 for the inferior or superior reader respectively. All quotients are multiplied by 100 to eliminate decimal points.

TABLE II

READING AGE NORMS

T Score	Reading Age	T Score	Reading Age	T Score	Reading Age	T Score	Reading Age
21	67	41	124	61	181	81	238
22	70	42	127	62	184	82	240
23	73	43	130	63	186	83	243
24	76	44	133	64	189	84	246
25	79	45	135	65	192	85	249
26	82	46	138	66	195	86	252
27	84	47	141	67	198	87	255
28	87	48	144	68	201	88	257
29	90	49	147	69	203	89	260
30	93	50	150	70	206	90	263
31	96	51	152	71	209	91	266
32	99	52	155	72	212	92	269
33	101	53	158	73	215	93	272
34	104	54	161	74	218	94	275
35	107	55	164	75	220	95	278
36	110	56	167	76	223	96	281
37	113	57	169	77	226	97	284
38	116	58	172	78	229	98	287
39	118	59	175	79	232	99	290
40	121	60	178	80	235	100	293

HOW TO INTERPRET T SCORES

The unit proposed above for measuring reading ability or any mental ability has been called T. The number of

TABLE 12

SAMPLE SCORE SHEET

School No. *11*, Grade *IV B*. Teacher *Miss X*. Date *June 15, 1920*.

I Chron. Age in Mos.	II Pupils' Names	III T Score	IV Reading Age	V Reading Quotient
124	*Adams, Sam*	*27*	*84*	*68*
120	*Baker, Mary*	*52*	*155*	*129*
148	*Davis, Geo.*	*40*	*121*	*82*
134	*Evans, Asa*	*46*	*138*	*103*
Mean		*41.3*	Mean	*95.5*
Grade Norm		*41.8*	Norm	*100*

questions correct is not a satisfactory unit for measuring reading ability because the difference in difficulty between 33 and 34 questions correct may be greater or less than between 12 and 13 questions correct. The difference in difficulty between 33 T and 34 T always equals that between 12 T and 13 T.

Again, T scores make possible such statements as the following: Any pupil whose T score is 50 has an ability which equals the average ability of all twelve-year-old children in the nation. Any pupil whose T score is 70 has an ability which is 20 T (or 2 sigmas) above the average ability of twelve-year-olds. Any pupil whose T score is 35 is 15 T (or 1.5 sigmas) below the average ability of twelve-year-olds.

Again, T scores may be interpreted thus:

A T Score of	Is Exceeded by the Following Per Cent of 12-year-olds	A T Score of	Is Exceeded by the Following Per Cent of 12-year-olds
25	99	55	31
30	98	60	16
35	93	65	7
40	84	70	2
45	69	75	1
50	50	80	0.1

HOW TO INTERPRET READING QUOTIENTS

How to interpret a Reading Quotient for a pupil or the mean Reading Quotient for a class is shown in the third column below. The second column gives the percentage of 5,000 pupils in grades II through VIII who had the Reading Quotients indicated in the first column.

Reading Quotient	Per Cent of Pupils	Interpretation
Below 55	0.1	
55 to 65	2.4	Exceptionally inferior
65 to 75	8.0	Very inferior
75 to 85	13.5	Inferior
85 to 95	18.0	Low average
95 to 105	19.6	Average
105 to 115	17.0	High average
115 to 125	10.5	Superior
125 to 135	5.8	Very superior
135 to 145	3.5	Exceptionally superior
145 up	1.5	

There are at least two useful ways of expressing the reading ability of a pupil (or of a class). How well a pupil reads is shown first by a comparison of his reading score with the norm for his grade. The use of this method alone encourages the school to retard pupils chronologically either unconsciously or consciously in order to give an appearance of high efficiency. Given a sufficient percentage of over-ageness or under-ageness, almost any school can appear efficient or inefficient respectively when compared with grade standards.

How well a pupil reads is shown, second, by a comparison of his reading score with the norm for his age. This is just what is expressed by the Reading Quotient. Showing as it does what the school has accomplished for the pupil by a given age rather than a given grade, the Reading Quotient has very great value, because the school cannot raise the Reading Quotient above 100 nor depress it below 100 by creating undue chronological retardation or acceleration respectively.

SUGGESTIONS FOR ECONOMIZING TIME

When the examiner has some skill in statistical procedure and is interested in class or grade or age scores only, it is not necessary to convert each pupil's number-of-questions-correct into a T score. Instead it will be a great saving of time to make a frequency distribution showing the number of pupils making each number-of-questions-correct. The first column of the frequency distribution can then be converted into T scores and the mean computed.

Again, when the examiner is interested only in reading ages and Reading Quotients time can be saved by copying column I of Table 9 into Table 11. This will save one step in the process of converting the number-of-questions-correct into a reading age.

Reading Achievement of Various Cities

At the End of	III	IV	V	VI	VII	VIII	IX	X	XI	XII
10 Miscellaneous Cities	32.8	40.9	46.3	51.7	58.0	59.8	60.7	62.5	64.3	67.0
33 Wisconsin Cities..	38.2	40.9	47.2	52.6	55.3	58.0				
18 Indiana Cities.....	40.0	49.9	58.9	67.0	68.8	71.5				
Louisville, Ky........		39.1	43.6	51.7	59.8	60.7				
New York City, N. Y.	36.5	41.0	47.5	51.5	55.8	58.4				
Paterson, N. J.......		35.5	40.9	49.0	51.7	53.5				
San Francisco, Cal...	37.3	45.4	53.5	52.6	58.9	62.5				
St. Paul, Minn.......	39.1	41.8	46.3	53.5	58.0	62.5				

ACCURACY OF NORMS

It is impossible to determine exactly how many pupils are included in the grade norms which accompany this test. An extremely conservative estimate is attempted in Table 10. The age norms given in Table 11 were read from a smoothed curve based upon the following number of pupils:

Age	7-8	8-9	9-10	10-11	11-12	12-13	13-14	14-15	15-16	16-17	Adult
Pupils	128	635	1275	1462	1561	1833	1662	1112	430	65	300

3. ***Convert the Initial Reading Score into a Reading Age.***—The foregoing *Directions for Using Thorndike-McCall Reading Scale* shows how to convert reading scores

into reading ages. The reading ages for our five pupils are shown in line 3 of Table 8.

The determination of reading age on the Thorndike-McCall Reading Scale is a simple matter because this test has age norms. In case the teacher is using some reading or other test which has grade norms and not age norms, the grade norms may be transmuted into age norms by means of a technique described in the preceding chapter.

One important function of this initial inventory of reading is to prevent re-teaching of abilities which have already been taught. Ayres and others have estimated the tremendous financial cost to the public of the 33 per cent of retardation in the schools. Someone has computed that $40,000,000 are spent annually re-teaching pupils. No one has been willing to estimate the loss to the retarded pupils.

Unfortunately the real retardation and its cost have been little studied. Retardation studies have called pupils retarded who were not retarded and overlooked retardation which was really present. This is still another fallacy which has resulted from a superficial study of surface appearances and one more argument for the use of educational tests to increase visibility for the really significant factors. Most pupils who are chronologically retarded are not educationally retarded at all. The only true cases of retardation are pupils who are kept below the grade for which they are fitted by educational age. Most of the chronologically retarded are where they belong educationally or, to be more exact, they are usually a little accelerated. It is the chronological accelerates who are usually most retarded educationally. Thus educational measurement justifies the rather queer conclusion that chronological retardation tends to mean educational acceleration. Contrary to usual thinking, the chief cost of re-teaching occurs with the latter rather than the former group of pupils. It is the chronological accelerates who are educationally retarded and who are re-taught something they already know. The chronologically retarded are, on the whole, re-taught something which they

failed to learn from one teaching. Of course, re-teaching these mentally inferior pupils is costly, but in the long run it is probably less expensive than to permit them to proceed without adequately mastering the prerequisites. The function, then, of the initial inventory is to prevent the cost to pupil and public of re-teaching what has really been learned.

A second function of the initial inventory is to avoid premature teaching. We have already seen how pupils are frequently started on a phase of the curriculum which, in the light of their measured capacity to learn, is too difficult for them. We saw again that pupils are frequently required to learn a portion of the curriculum before they have learned certain prerequisites in the hierarchy. The initial inventory will not only prevent such premature teaching in general, but will definitely point out for the guidance of both learner and teacher just where the pupil is most deficient and hence where he most needs help. Two of the great wastes in education are due to re-teaching or premature teaching. An adequate initial inventory will prevent both. Says Foote, "When pupils and teachers know where they are and where they are to go there is reason to believe that the journey will be accomplished; otherwise it is very doubtful." It is the function of the initial inventory to show pupils and teachers *where they are.*

4. ***Determine the Reading Quotient.***—The previously quoted *Directions for Using Thorndike-McCall Reading Scale* describes how to compute and interpret Reading Quotients. A similar procedure and interpretation applies to Arithmetic Quotients, Spelling Quotients and the like.

5. ***Determine the Initial Mental Age.***—It is recommended that the teacher use, for determining the mental age of each pupil, the Stanford Revision of the Binet-Simon Intelligence Scale sold by Houghton Mifflin Company, New York City, or in case the school is in a decidedly foreign neighborhood, the Pintner-Paterson Scale of Performance Tests sold by D. Appleton & Company, New York, or the National Intelligence Test sold by World Book Com-

pany, Yonkers-on-Hudson, N. Y. The first two tests require fairly expensive equipment and a trained examiner. The National Intelligence Scale is relatively inexpensive, can be given to all the pupils in a class or school at one time, and yields fairly satisfactory results even when used by a complete novice. Most teachers should use the last test. For grades below the third the teacher may use either Haggerty's Intelligence Test, Delta I, World Book Company, Yonkers-on-Hudson, N. Y., or Pressey's Primer Scale, University of Indiana, Bloomington, Indiana, or Dearborn's Intelligence Test for these grades, Lippincott Company, Philadelphia, Pa.

The method of scoring the first two tests yields a mental age score directly. The score on the National Intelligence Test must be converted into a mental age just as the reading scores were converted into reading ages. Age standards are sent with this test.

The initial mental age of pupil A is 121. It, together with the mental ages for the other 4 pupils, is shown in line 5 of Table 8.

6. ***Determine the Intelligence Quotient.***—A pupil's Intelligence Quotient (I.Q.) is the quotient of his mental age divided by his chronological age. Thus pupil A's I.Q. is 121 divided by 121, i. e., 100. The I.Q. of pupil B is 150 divided by 123, i. e., 122. The I.Q. for each of the other three pupils is shown in line 6 of Table 8.

A knowledge of a pupil's I.Q. should be of very great value to any teacher of any subject, for the size of a pupil's I.Q. is an index of his general mental brightness or mental alertness. As Terman points out the most important fact about a pupil, next to character, is his I.Q. The significance of I.Q.'s of varying sizes is brought out below:

Above 140	Genius or near genius.
120 — 140	Very superior intelligence.
110 — 120	Superior intelligence.
90 — 110	Normal or average intelligence.
80 — 90	Dullness.
70 — 80	Borderline deficiency, sometimes feeblemindedness.
Below 70	Definitely feebleminded.

These determinations of reading age and Reading Quotient, and mental age and Intelligence Quotient not only furnish valuable teaching guides but also provide the basis for educational guidance through a knowledge of a pupil's capacity to profit by general education and pursue particular subjects.

General Capacity to Learn.—One problem in education is to locate the educational objectives. Another is to locate somebody who has the capacity to attain these objectives —to find somebody who is educable. Pigs, sheep, cows, horses, dogs, and other domesticated animals have widely varying capacities to learn. While the percentage of illiteracy is high, these animals have a more or less definite curriculum and are taught certain lessons by their owners. It is only human beings, however, who are considered to have enough capacity to learn to make systematic, prolonged education profitable.

But the technique of diagnosing capacity to learn does not end with the classification of an animal as a human animal. The range of capacity to learn among humans is greater than the difference between humans in general and dogs in general. The overlapping of capacity is so great that a considerable per cent of humans have a capacity to learn which is inferior to the geniuses among dogs, cats, monkeys, and other much reviled creatures.

In brief the procedure in diagnosing learning capacity is to measure what the individual can learn or has learned. The former method is to place a child, say, in a novel situation and score how well he learns. The latter method is to assume that he has since birth been in a learning situation and to measure how much he has learned.

The first measurements of capacity to learn were simple unstandardized observation of children by parents and neighbors. These measurements were inevitably inaccurate because of numerous constant errors such as parental vanity, neighborly jealousies, absence of constant or fair standards of estimate as well as other more subtle errors.

These subjective measurements were probably more accurate a generation ago than at present, because numerous progeny facilitated the development of surer standards of measurement.

As a result of parental measurements the extremely stupid children were kept at home while those with a greater capacity were sent to school. Many of the stupidest ones were committed to institutions for the feebleminded, while the stupider ones were sent to special schools, classes, tutors, and the like.

When children are dumped into the hopper of the educational mill, they enter a great and more accurate selective machine. Every stage of education from the kindergarten through the university is engaged in the process of selection. In a very real sense our schools are as much selective as educative agencies. Every teacher takes her toll, though gallantry forbids saying that this is a case of the devil takes the hindmost. There is a miller whose water mill on the Tennessee River grinds the corn for the farmers for miles around. The miller always takes his toll from the best bag of corn. Teachers are more generous; they take toll from the children whose capacity to learn is least. Each year the ranks of this grand army of children grows thinner and thinner. The Ph.D. or the equivalent is the educator's reward for the students who have been able enough or clever enough to escape the clutch of all the teachers.

More and more educational selection is becoming an important function of the school. Children are being committed to institutions for the feebleminded. This is frequently construed as a stigma upon both children and parents. Private schools deny entrance to children whose learning capacity is judged to be below a certain standard. Public schools are sending pupils to special classes for the mentally slow. Stupid pupils are denied promotion. Certain public schools group pupils within each grade according to learning capacity. Other public schools refuse admission to any whose learning capacity is not unusually

great. Some countries, recognizing that its greatest asset is its children of genius and that these geniuses belong to the community rather than to particular parents, are selecting these children for a special education.

When matters of such critical importance to the individual are at stake a democracy will not long tolerate a system of educational selection which does not utilize the most thoroughly scientific, impartial, impersonal, and rigidly standardized technique possible. Standardized educational and psychological tests, inaccurate though they may be, are rapidly becoming recognized as the best means for educational selection. It is but a question of time until they supplant the traditional selective mechanism of home and school.

The use of tests for selective purposes unquestionably demonstrated their worth during the recent war. Influenced by this demonstration, Columbia College adopted the Thorndike College Entrance Intelligence Tests as a part of the machinery for selecting its students. Other colleges and many lower schools admit students on the basis of standardized measurements, while commitment of children to institutions for the feebleminded has long been a function of the Binet-Simon Intelligence Test.

Lack of suitable group tests has delayed the movement toward the adoption of standard tests as a means for determining learning capacity. This lack is now being rapidly supplied. With the completion of the National Intelligence Tests, we have an instrument whose use should be universal for numerous purposes, particularly for discovering instances of unusual capacity. The last chapter in this book lists many other tests which will be found helpful.

Psychologists are now able to tell with considerable accuracy whether a child possesses an I.Q. which will ever make it possible for him to do the work of a particular school or institution or grade in a school. Further, they are able to determine whether a child's mental age is now sufficient to learn the work of a particular grade. Terman's

experience leads him to the conclusion that the 60 I.Q. pupil will not be able to do work beyond Grade III or IV. The 70 I.Q. child will not be able to do work beyond Grade V or VI. The 80 I.Q. will reach his limit about Grade VII. The 90 I.Q. pupil may by dint of much persistence go through high school. E.Q.'s of 60, 70, 80 and 90 for pupils whose educational opportunities have been normal may be interpreted like similar I.Q.'s. Even the attainment listed above cannot be reached until the mental age or educational age has sufficiently developed and this means considerable chronological retardation.

Capacity to Learn a Particular Subject.—Mary inquires of her teacher whether it is wise, considering the strength of her purpose and the extent of her capacity, for her to study high-school mathematics. John walks into the principal's office and innocently asks such a simple question as: "Do you consider it best for me to study a foreign language?" Similar questions may be, though they less frequently are, asked about elementary school abilities. The determination of the capacity to learn in a particular subject is more difficult than to determine capacity in general. Three methods have been employed:

1. A measurement of the pupil's previous achievement and rate of progress in a subject which continues into the future. This method measures capacity inclusive of purpose. Reading, writing, arithmetic, spelling, composition, etc., are subjects which are more or less continuous through the elementary school. Ability in these subjects can be measured by group tests from about Grade III through Grade VIII. Hence within these ranges pupils may be advised for the future on the basis of their Reading Quotients, Spelling Quotients, etc., in these same subjects in the past. In the high school the student may be assigned tentatively to a certain study and asked to pursue it for a brief period, after which he is measured. His prospects may be estimated from his recent progress.

2. A measurement of the mental abilities upon which

success in the subject to be pursued depends. This method includes purposes only indirectly. Success in high school mathematics, for example, depends upon the possession of certain special mental abilities. Dr. Rogers [1] attempted to analyze the mental elements in mathematical ability and to construct tests for their measurement. If the analysis is correct, and if the tests really measure the subordinate abilities discovered, and if a pupil is shown by the tests to be of the requisite abilities, then we can make an accurate prognosis concerning his prospects in, say, geometry.

3. A measurement of general intelligence. This method of making a prognostic diagnosis makes no pretense to a precise analysis of the abilities required. It is known that scholastic achievement, whether in the elementary school, high school, or college, is closely correlated with general mental ability. Hence a knowledge of the pupil's intelligence enables one to make a fairly accurate prognosis.

The whether-to-pursue-a-subject question is intimately allied with the when-to-begin-a-subject question. The former must consider whether the value, purpose, and capacity are now or ever will be adequate. The latter assumes that purpose and capacity will be adequate when experience and training have accumulated and maturity is far enough advanced. The answer to both questions requires a knowledge of the limit of purpose plus capacity which must be exceeded to bring success.

Our ignorance of this limit is not absolute. Educators and psychologists are now able to set crude limits, but they are very crude. A short time ago a teacher who is respected and admired for her knowledge of children and her skill in teaching them confessed that she did not know, for example, such a simple fact as the time when a child has both a purpose and capacity for reading. She proposed to set about determining this time.

But the definition of these limits is not as simple a prob-

[1] Agnes L. Rogers, *Experimental Tests of Mathematical Ability and Their Prognostic Value;* Bureau of Publication, Teachers College, Columbia University, N. Y. C., 1918.

lem as she conceived it to be. It is not simple because ability to read does not suddenly and all at once leap into existence. It gradually grows. By dint of enough effort reading could be taught much earlier than it is. Furthermore, ability for other activities is developing at the same time, so it becomes necessary to decide which of many abilities reach white heat first. Again, individuals differ in their rate of development. Finally, the whole question is complicated by the number of years of training required to reach the desired goal, and the age at which the pupil's daily needs require the attainment of a given goal.

Unfortunately lack of knowledge of this whole question has resulted not so much in a strenuous desire for scientific study as in a sort of false sentimentalism. Many pupils are working far below their possible efficiency because of a fear that their neurones will be injured—that their minds will be strained! If minds were so easily strained, infants would be mentally stunted for life. The neurones are never again subjected to such a strain as when both parents and relatives are trying to wrest from the neural mechanism a faint approximation to "ma-ma," "pa-pa." Until he is entirely ready to gratify these parents, uncles, and aunts, the baby sweetly and tolerantly smiles at their silly behavior, and continues to protect his tender neurones. Nature has the same excellent device for protecting pupils. When the pupil is presented with a problem beyond his capacity Nature automatically cuts the circuit and nothing happens.

7. ***Determine the Initial Accomplishment Quotient.*** —The customary procedure of comparing the class score with the grade norm is not as useful for measuring past efficiency as an Accomplishment Quotient, because the customary procedure totally disregards the intellectual calibre of the class. The Accomplishment Quotient is found by dividing reading age by mental age. Pupil A's reading age in Table 8 is 121 and his mental age is 121. Hence his Accomplishment Quotient for comprehension while reading silently is 121 divided by 121, i. e., 100. The Accomplish-

ment Quotients for the other pupils are shown in line 7 of Table 8.

The Accomplishment Quotient is the most exact present-day measure of the efficiency of study, instruction, and supervision; it is the only just basis for reporting to parents and for judging pupils; and it is the best index of what pupils need special attention and spurring, of what pupils need restraining perhaps, and of what pupils need to be "let alone."

It is a common occurrence for pupils of low general intelligence to be placed in their chronological group and told to keep up with the class or be publicly stigmatized, officially denied promotion, and corporally punished at home. The Accomplishment Quotient holds out a promise of much-needed justice to such pupils. It asks the pupil to progress at a rate which is proportional to the mental capacity with which nature endowed him.

It is a common occurrence for a pupil to be urged forward when for the sake of his health, possibly, he should be restrained, and for another pupil to be restrained who should be prodded. Pupil C is at the bottom of the class. In the conventional school he would be judged highly inefficient and all the weight of the school would be brought to bear upon him. Pupil B on the other hand would be praised for his excellent work, and would be given high marks. As a matter of fact pupil C has made a better use of his capacities than any other pupil in the class and pupil B is about the most inefficient. Because of improper instruction pupil B has not "kept the traces taut."

It is a common occurrence for teachers to be assigned to a class of stupid pupils and to be expected to produce standard results. Such a policy is very unfair. The Accomplishment Quotient promises protection against such injustice.

Sometimes a teacher is assigned to a group of pupils of normal intelligence which has been improperly taught for

several years. It is unfair to expect a teacher to overcome in one year several years of neglect. The mean Accomplishment Quotient for the pupils in Table 8 shows that this class has made the progress in the past of which it was capable and thus shows both the teacher and her supervisor that there is no initial handicap.

That pupil or class which has an Accomplishment Quotient of 100 has made satisfactory progress. Consistent with health and the need for developing other abilities, the teacher should aim to keep the Accomplishment Quotient for reading as much above 100 as possible. As a rule the teacher's first task will be to assist the young gifted children. Due to their being classified below their proper level and to neglect in general their Accomplishment Quotients are usually lower than that of any other pupils.

In the present stage of mental measurement even the Accomplishment Quotient is crude. It is crude because both the numerator and denominator of the formula are imperfect. There is an immense segment of abilities and purposes which should be in an ideal curriculum and which are not included in educational age, as at present determined. Educational age, as determined at present, is satisfactory only to the extent that it is representative of the total abilities and purposes which should be the objectives of instruction.

Similarly, the denominator is inadequate. There are native equipments other than those usually measured by intelligence tests which condition school progress. Teachers realize that pupils are not all intelligence or stupidity. Some are intelligent and lazy while some are stupid and industrious. To the extent that industry, persistence, conscientiousness, etc., are results of instruction they belong in the numerator while to the extent they are native equipment they belong in the denominator. Thus much elaborate research is necessary before a thoroughly satisfactory Efficiency Quotient can be computed.

II. Method of Diagnosis

Method and Function of Diagnosis.—The tree Igdrasil pictures the problem of diagnosis. Carlyle describes Igdrasil as the ash-tree of existence which has its roots deep-down into the kingdom of hela, whose trunk reaches heaven-high, and whose boughs spread over the whole universe, a tree which is the past, present, and future, and what was done, is doing, and will be done. A central ability or purpose in a pupil is a miniature Igdrasil. Its roots reach deep-down into the educational conditions of early days, and its boughs spread through all his mental life; it shows the past, the present, and the future, and what was done, is doing, and will be done.

One bough of reading ability reaches into reasoning problems in arithmetic. The initial inventory reveals that the ability to solve written problems is defective. It then becomes the business of diagnosis to locate the cause, and the cause of the cause, and the cause of the cause of the cause, and so on back to the teaching unit. In sum it becomes the task of diagnosis to trace a miniature Igdrasil from leaf to root. In the illustration it is the task of diagnosis to discover that the cause of inability to solve problems is a defective reading ability, and that the cause of a defective reading ability is an inadequate vocabulary and so on. Thus the method of diagnosis is to trace abilities to their roots by means of standardized tests in order to discover just which ability or element of it exists out of standard proportions. This is the method of locating the underlying causes of defects.

The function of diagnosis is to guide corrective measures. There is an inscription upon the monument which commemorates the arrival of the first white man at the Cumberland River in Central Tennessee. The inscription is to his wife and reads thus: "She shed a leading light along his path of destiny." Diagnosis is the veritable wife of remedial instruction. Without its guidance corrective in-

struction is absolutely "hit or miss," with but one chance to hit and several million chances to miss.

There are an enormous number of diagnoses being made in our schools daily. Some of these diagnostic measurements are vague and penumbral and some are quite exact. Every increase in the accuracy of the diagnostic measurements means an increase in the percentage of hits. To make these diagnoses accurate requires time, but so does teaching. Many teachers do not realize that a large per cent of their pupils have not advanced one iota as a result of a year's teaching in, say, fundamentals of arithmetic. Diagnosis would mean a net saving of time.

Diagnostic Methods: Introspection by Pupil.—This method is so obvious and is so frequently employed that it needs neither discussion nor illustration. Pupils frequently know not only the exact location of their difficulty but the cause of the difficulty as well. When the pupil is able to diagnose his own difficulty it is a waste of time and effort for the teacher to resort to the more elaborate methods yet to be described. Even when the pupil does not thoroughly understand his difficulty a conversation with him may give the more experienced teacher sufficient data to make a diagnosis.

Diagnostic Methods: Observation of Normal Work.—The commonest method of diagnosis is to get some hint from the behavior of the subject being diagnosed. When school was not in session we three brothers worked in the mines with our father. He was particularly expert in diagnosing the condition of the rock under which we worked and in detecting the imminence of danger. For this reason he was always assigned to the dangerous task of removing the last coal which supported the overhanging rock. As more and more of the coal was removed the weight of the millions of tons of rock slowly settled upon the frail wooden timbers. They would become taut like the strings of a violin, so that flying splinters caused by the pressure made a sort of music. Occasionally a timber would break

with a sharp sound like the crack of a rifle. Through it all father worked as though unhearing. Perhaps a week later he would say: "Get your tools, boys, and get out as fast as you can." We would go a short distance to a place of safety, lie down behind a car so as not to be struck by loose objects blown by the wind of the fall, and listen to the snapping of the props and the grinding of the mountain. As we grew older we, too, learned to interpret hints given by the rock. Here as with wild things in the woods it was diagnosis or death and diagnosis from subtle behavior hints.

The teacher watches a pupil read who is having difficulty with reading. She observes that his eyes do not have three or four evenly-spaced brief fixations per line, but move forward, then jump back again, and act in a generally irregular fashion. Observation of this behavior aids the trained teacher to make a diagnosis of the difficulty. Another pupil is rarely able to complete an assignment in history. By observing his study the teacher notes that while reading his lesson he screws up his face, shakes his head, moves his lips, and tugs at his hair. This, too, is a hint to the perceiving teacher. Another pupil is very slow at figures. The discerning teacher may construct a trial diagnosis by noting that he is counting with his fingers, toes, or tongue, and whispering as he adds: "Seven and six make thirteen, and thirteen and eight make twenty-one." Another pupil is having trouble with division of fractions. An examination of his written work may reveal that the source of his difficulty is failure to invert the divisor. Thus accurate, detailed, trained, and experienced observation of pupils in the process of normal work is one method of discovering the data upon which to base a diagnosis and prescribe corrective measures.

Courtis [2] has listed some arithmetical defects discovered by this diagnostic method. Along with the defects he gives

[2] S. A. Courtis, *Teacher's Manual for the Standard Practice Tests;* World Book Co., Yonkers-on-Hudson, N. Y., copyrighted 1915. Used by permission of the publishers.

an excellent statement of the underlying causes and suggests corrections.

"1. Child's movements very slow and deliberate, but steady.

"2. Child's movements rapid but variable. Adding accompanied by general restlessness, sighs, frowns, and other symptoms of nervous strain.

"3. Child's progress up the column irregular; rapid advance at times with hesitation, or waits, at regular or irregular intervals. Often gives up and commences a column again.

"4. Child stops to count on fingers, or by making dots with pencil, or to work out in its head the addition of certain figures.

"5. Child adds each first column correctly, but misses often on second and third columns.

"6. Child's time per example increases steadily or irregularly; particularly after two or three minutes' work; i. e., 15 seconds each for first five examples, 17 seconds each for the next five, 23 seconds for next two, 45 seconds for the next example, etc.

"7. Child's habits apparently good and work steady, but answers wrong."

The diagnosis and correctives follow:

"1. Slow movements may be due either to bad habits of work or to slow nerve action. In the latter case, the difficulty will prove very hard to control. It is almost certain that no amount of training will ever alter the nerve structure and so remedy the fundamental cause. But in all such cases much can be done to generate ideals of speed, to help the child to eliminate waste motions, and to hold himself up to his best rate.

"In any case the procedure would be as follows: Ask the child to add the first example alone so that you may time him. Give him the signal when to start and let him signal when he has finished. Let him make several trials of the same example to make sure that he does not improve under practice. The teacher should then give the child the

watch and let him *time the teacher in working the same example.* Comment on difference in child's and teacher's times. Then have the child write in small figures all the partial sums, as shown in the illustration. The teacher should again time the child, letting him *read* to himself the partial sums as rapidly as he can. This will, of course, give the minimum time in which the child could possibly add the example. The time records of a child with true defective motor control will show slight improvement, if any, even with such aid, and probably the only procedure to follow in such cases is to lower the standard to correspond. Where there is a marked difference in time between the original and this last performance, the child will get, for the first time in its life, perhaps, a perfectly clear *conception* of what working at standard speed really means, as well as the *sensation* of really working at that speed. The teacher and child should then practice the same example over and over until the child can *without the crutches* add it at the standard rate. Now the teacher can give him the whole test again, urging him to work at his best speed and comparing his results with the first result. The improvement made by ten minutes of this kind of work enables the teacher to say that a proper amount of similar study would produce the changes desired.

———
30 15
46
26 9
41
22 8
97
13
60
7
61

" 'But,' some teacher will say, 'will the child not learn the example by heart?' This is precisely what is desired. A perfect adder has learned so many examples 'by heart' that it is impossible to make up any arrangement of figures that will be in any way new to him. The child in the same way needs to perfect his control over *each* example until he finally attains to mastery over *all.*

"2. If the child gives evidence of nervous strain, check his speed, teach him to relax and to work easily and quietly. Get good habits of work first, then bring up speed and accuracy by degrees. The nervousness of a child is usually caused by social conditions, physical health, or temperamental bias. In any event it is difficult to control. Look out for a large fatigue factor in nervous children.

"3. Irregular speed up the column may be due to either of two factors: lack of control of attention, or lack of knowledge of the combinations. The latter factor will be discussed in the following paragraph (4). Attention will be considered here.

"There is a limit to the length of time that a person can carry on any mental activity continuously. As time goes on, the mind tends to respond more and more readily to *any* new mental stimulus than it does to the old. The mind 'wanders' as it is said. The attention span for many children is six additions, for some only three or four, for others eight, or ten, and so on. That is, a child whose attention span is limited to six figures may add rapidly, smoothly, and accurately, for the first five figures in the column, giving its attention wholly to the work. As the limit of its attention span is reached, however, it becomes increasingly difficult for it to concentrate its attention. The child suddenly becomes conscious of its own physical fatigue, of the sights and sounds around it. The mind balks at the next addition; it may be a simple combination, as adding 2 to the partial sum, 27, held in mind. It finally becomes imperative that the child momentarily interrupt its adding activity and attend to something else. If this is done for a small fraction of a second, the mind clears and the adding activity will go on smoothly for a second group of six figures, when the inattention must be repeated.

"It should be evident that these periods of inattention are critical periods. If the sum to be held in mind is 27, there is great danger that it will be remembered as 17, 37, 26, or some other amount, as the attention returns to the work of adding. The child must, therefore, learn to 'bridge' its attention spans successfully. It must learn to recognize the critical period when it occurs, consciously to divert its attention while giving its mind to remembering accurately the sum of the figures already added. This is probably best done by mechanically repeating to one's self mentally, 'twenty-seven, twenty-seven, twenty-seven,' or whatever the sum may be, during the whole interval of inattention. Little is known about the different methods of bridging the attention spans and it may well be that other methods would

prove more effective. The use of the device suggested above, however, is common.

"Giving up in the middle of a column and commencing again at the beginning is almost a certain symptom of lack of control of the attention. On the other hand, mere inaccuracy of addition (as 27 plus 2 equals 28) may be due to lack of control over the combinations. If the errors occur at more or less regular points in a column, and if, further, the combinations missed vary slightly when the column is re-added, the difficulty is pretty sure to be one of attention and not one of knowledge.

"4. Hesitation in adding the next figure, when not due to attention, is usually due to lack of control of the fundamental combinations. In such cases, however, the hesitation or mistakes are usually repeated *at the same point* on subsequent additions. The teacher should understand that it 'takes time to make mistakes,' and whenever a lengthening of the time interval occurs, it is a symptom of a difficulty which must be found and remedied.

"In this case the remedy is *not* a study of the separate combinations. It has been proved [3] that for most children time spent in study of the tables is waste effort; that the abilities generated are specific and do not transfer. A child may know 6 plus 9 perfectly, and yet not be able to add 9 to 26 in column addition except by counting on its fingers. *The combinations must be learned, of course, but they should be learned by practicing column addition.* Follow the method outlined in paragraph (1) above, having the column added over and over again until both standard speed and absolute accuracy have been attained.

"5. The sums of a child who is unable to remember the numbers to be carried, but whose work is otherwise perfect, will usually have the first column added correctly, as well as all single columns. Unfortunately, however, inability to carry correctly is usually a fault of children with weak memories for partial sums in the column. It is well, therefore, to test the carrying habits of any child that is inac-

[3] See Bulletin No. 2, Department of Coöperative Research, Courtis Standard Tests, 82 Eliot St., Detroit, Mich. Price 15 cents. See also, *Journal of Educational Psychology*, September, 1914.

curate. Many children do not add the number carried until the end of the next column; it should, of course, be added to the first figure in the column. If necessary the number to be carried should be emphasized as by saying, when the sum of a column is 27, 'carry 2' to one's self as the 7 is written. This is again a time-consuming device which should be adopted only as a last resort. The carrying should be an automatic, unconscious operation. Repeated practice on a few examples until the same become so perfectly familiar that a child's whole attention may be given to establishing correct habits of carrying will prove beneficial.

"6. Marked increases in the times required for the successive examples of a test are an indication of a fatigue factor in the control of the attention. Some children are unable to carry on continuously a single activity, as adding, through even a four-minute time interval without a very great loss in power. Two courses are open to the teacher, one or the other of which is sometimes effective: one is to determine the exact length of the interval at which the child can work efficiently, and then try to extend the interval slightly each day; the other is to set the child at work on very long and very hard examples, and to lengthen the time intervals to fifteen or twenty minutes' continuous work. Difficulties of this type are hard to remedy."

Diagnostic Methods: Oral Tracing of Process.—There are difficulties the underlying causes of which would never come to light from an introspective inquiry on the part of the pupil or from mere observation of the pupil's normal work. The purpose of the diagnostic process is, of course, to induce the pupil to commit some overt act which will reveal the invisible causes of his visible defect. When neither his ordinary actions nor his written work offers a suggestion it is well to have the pupil go through the process orally. When I fail to make the class in educational measurement understand the computation of, say, the median, I find it advantageous to ask one of the students who is having

trouble to come to the blackboard and compute a median orally for the class. The cause of the difficulty is thus quickly found.

Uhl used the oral-tracing method to discover the mental processes through which pupils go in adding and subtracting. The old phrase: "Beat the devil around the stump" accurately describes how some pupils work. To quote Uhl: [4]

"The findings as to methods employed by pupils in 'difficult' combinations is both interesting and significant. The following methods were found in the work of pupils who were tried out in the manner just described. A fourth-grade boy showed by slow work that the combination 9 — 7 — 5 was difficult for him. When questioned, he showed that he used a common form of 'breaking-up' the larger digits. In working the problem, he said to himself: '9 + 2 + 2 + 2 + 1 = 16 and 21.' This shows that the 9 — 7 combination was not known, but that the 16 — 5 combination was, inasmuch as he arrived at '21' directly after having combined the other two numbers. Another boy of the same grade showed the same type of difficulty in a more pronounced form. He added 8, 6, and 0 as follows: 'First take 4, then take 2, then add 8, and 4 makes 12, and 2 makes 14.' In adding 9, 7, and 5 he said: '9 and 3 is 12 and 4 is 16 and 2 — 18; and 2 — 20; and 1 — 21.' He broke into parts even so easy a problem as 3 + 4 + 9, adding 9 + 3 + 2 + 2 = 16.

"A pupil from the fifth grade presented a quite different method of adding. In adding 4, 9, and 6 she explained: 'Take the 6, then add 3 out of the 4. Then 9 and 9 are 18, and 1 are 19.' Other problems were worked out similarly: one containing 3, 9, and 8 was solved as follows: '8 and 8 are 16 and 3 are 19 and 1 are 20'; 5, 6, and 9 as follows: '6, 7, 8, 9, and 9 are 18 and 2 are 20.' This tendency to build up combinations of 8's or 9's continued in the case of another problem: 6, 5, and 8 were added thus: '6, 7, 8, and 8 are 16 and 3 are 19.' Probably her first problem was worked similarly, but I had to have her dictate her

[4] W. L. Uhl, "The Use of Standardized Materials in Arithmetic for Diagnosing Pupils' Methods of Work"; *Elementary School Journal*, November, 1917.

method twice before I understood; she then gave it as quoted.

"Methods which are quite as clumsy are found in the case of subtraction. One boy of the fifth grade was found to build up his subtrahend in the case of many problems. For example, in subtracting 8 from 37, he increased his subtrahend to 10, then obtained 27, and finally added 2 to 27 to compensate for the addition of 2 to 8. Likewise, in subtracting 7 from 30, he added 3 to 7 and proceeded as before. This boy knew certain combinations very well, but did problems containing other combinations by a method much harder than the correct one.

"Even greater resourcefulness was shown by a fifth-grade boy who found the differences between some numbers by first dividing, then noting the remainder or lack of one, then multiplying, and finally adding to or taking from the result as necessary. For example, in subtracting 9 from 44, he proceeded as follows: 'Nine goes into 44 five times and 1 less; 4 times 9 are 36, minus 1 equals 35.' That is, this boy knew certain multiplication combinations better than he did certain subtraction processes; therefore, he used multiplication, making adjustments either upward or downward as demanded by the problem."

Diagnostic Methods: Analysis of Test Results.—There are many tests specially designed not only to measure in the usual sense but to facilitate diagnosis. Monroe's Diagnostic Tests in Arithmetic is an illustration of such tests. Practically every standard test has some diagnostic value.

Using his Reading Scale Alpha 2, Thorndike made an unusually subtle analysis of pupil results to discover the causes for imperfect comprehension in reading. The following selected quotations [5] will increase anyone's respect for the mental process called *reading* and will show the problem a teacher faces who undertakes to teach or diagnose this complex ability.

[5] E. L. Thorndike, "Reading as Reasoning: A Study of Mistakes in Paragraph Reading"; *Journal of Educational Psychology,* June, 1917.

"It will be the aim of this article to show that reading is a very elaborate procedure, involving a weighing of each of many elements in a sentence, their organization in the proper relations one to another, the selection of certain of their connotations and the rejection of others, and the coöperation of many forces to determine final response. In fact we shall find that the act of answering simple questions about a simple paragraph . . . includes all the features characteristic of typical reasoning. . . .

"In correct reading (1) each word produces a correct meaning, (2) each such element of meaning is given a correct weight in comparison with the others, and (3) the resulting ideas are examined and validated to make sure that they satisfy the meaning set or adjustment or purpose for whose sake the reading was done. Reading may be wrong or inadequate (1) because of wrong connections with the words singly, (2) because of over-potency or under-potency of elements, or (3) because of failure to treat the ideas produced by the reading as provisional, and so to inspect and welcome or reject them as they appear. . . .

"In particular, the relational words, such as pronouns, conjunctions and prepositions, have meanings of many degrees of exactitude. They also vary in different individuals in the amount of force they exert. A pupil may know exactly what *though* means, but he may treat a sentence containing it much as he would treat the same sentence with *and* or *or* or *if* in the place of the *though.*

"The importance of the correct weighting of each element is less appreciated. It is very great, a very large percentage of the mistakes made being due to the over-potency of certain elements or the under-potency of others. . . .

"To make a long story short, inspection of the mistakes shows that the potency of any word or word group in a question may be far above or far below its proper amount in relation to the rest of the question. The same holds for any word or word group in the paragraph. Understanding a paragraph implies keeping these respective weights in proper proportion from the start or varying their proportions until they together evoke a response which satisfies the purpose of the reading.

"Understanding a paragraph is like solving a problem in mathematics. It consists in selecting the right elements of the situation and putting them together in the right relations, and also with the right amount of weight or influence or force for each. The mind is assailed as it were by every word in the paragraph. It must select, repress, soften, emphasize, correlate and organize, all under the influence of the right mental set or purpose or demand.

"Consider the complexity of the task in even a very simple case such as answering question 6 on paragraph D, in the case of children of grades 6, 7 and 8 who well understand the question itself.

"John had two brothers who were both tall. Their names were Will and Fred. John's sister, who was short, was named Mary. John liked Fred better than either of the others. All of these children except Will had red hair. He had brown hair.

"6. *Who had red hair?*

"The mind has to suppress a strong tendency for *Will had red hair* to act irrespective of the *except* which precedes it. It has to suppress a tendency for *all these children . . . had red hair* to act irrespective of the *except Will.* It has to suppress weaker tendencies for *John, Fred, Mary, John and Fred, Mary and Fred, Mary and Will, Mary, Fred and Will,* and every other combination that could be a '*who,*' to act irrespective of the satisfying of the requirement 'had red hair according to the paragraph.' It has to suppress tendencies for John and Will or brown and red to exchange places in memory, for irrelevant ideas like *nobody* or *brothers* or *children* to arise. That it has to suppress them is shown by the failures to do so which occur. The *Will had red hair* in fact causes one-fifth of children in grades 6, 7 and 8 to answer wrongly,[6] and about two-fifths of children in grades 3, 4 and 5. Insufficient potency of *except Will*[7] makes about one child in twenty in grades 6, 7 and 8 answer wrongly with 'all the children,' 'all,' or 'Will, Fred, Mary and John.'"

[6, 7] Some of these errors are due to essential ignorance of "except," though that should not be common in pupils of grade 6 or higher.

After completing a thorough analysis of results from tests of pupils' ability to solve arithmetic problems, Monroe diagnosed many of the errors as due to inability to read the problems, inability to calculate accurately, and inability to reason correctly, which are in turn due to still more fundamental causes. According to Monroe,[8] pupils' mental processes when reasoning incorrectly are fairly pictured by Adams'[9] description of how the canny Scotch pupils solved this freak of a problem: "If 7 and 2 make 10, what will 12 and 6 make?" The description follows:

"A look of dismay passed over the seventy-odd faces as this apparently meaningless question was read. Everybody knew that 7 and 2 didn't make 10, so that was nonsense. But even if it had been sense, what was the use of it? For everybody knew that 12 and 6 make 18—nobody needed the help of 7 and 2 to find that out. Nobody knew exactly how to treat this strange problem.

"Fat John Thomson, from the foot of the class, raised his hand, and when asked what he wanted, said:

"'Please, sir, what rule is it?'

"Mr. Leckie smiled as he answered:

"'You must find out for yourself, John; what rule do you think it is, now?'

"But John had nothing to say to such foolishness. 'What's the use of giving a fellow a count[10] and not telling him the rule?'—that's what John thought. But as it was a heinous sin in Standard VI (seventh grade) to have 'nothing on your slate,' John proceeded to put down various figures and dots, and then went on to divide and multiply them time about.

"He first multiplied 7 by 2 and got 14. Then, dividing by 10, he got 1 2/5. But he didn't like the look of this. He hated fractions. Besides, he knew from bitter experience that whenever he had fractions in his answer he was wrong.

"So he multiplied 14 by 10 this time, and got 140, which certainly looked much better, and caused less trouble.

"He thought that 12 ought to come out of 140; they both looked nice, easy, good-natured numbers. But when he found that the answer was 11 and 8 over, he knew that he had not yet hit upon the right tack; for remainders are just as fatal in answers as fractions. At least, that was John's experience.

"Accordingly, he rubbed out this false move into division, and fell back upon multiplication. When he had multiplied 140 by 12, he found the answer 1680, which seemed to him a fine, big, sensible sort of answer.

"Then he began to wonder whether division was going to work this time. As he proceeded to divide by 6, his eyes gleamed with triumph.

"'Six into 48, 8 an' nothin' over, — 2 — 8 — 0 an' no remainder. I've got it!'

[8] Walter S. Monroe, *Measuring the Results of Teaching*, pp. 154-172; Houghton Mifflin Co., N. Y., 1918.

[9] John Adams, *Exposition and Illustration in Teaching*, pp. 176-178.

[10] Scotch: Any kind of arithmetical exercise in school.

"Here poor John fell back in his seat, folded his arms, and waited patiently till his less fortunate fellows had finished.

"James[11] knew from the 'if' at the beginning of the question that it must be proportion; and since there were five terms, it must be compound proportion. That was all plain enough, so he started, following his rule:

" 'If 7 gives 10, what will 2 give?—less.'

"Then he put down

7 : 2 :: 10 :

" 'Then if 12 gives 10, what will 6 give?—again less.' So he put down this time

12 : 6

"Then he went on loyally to follow his rule: multiplied all the second and third terms together, and duly divided by the product of the first two terms. This gave the very unpromising answer 1 3/7.

"He did not at all see how 12 and 6 could make 1 3/7. But that wasn't his lookout. Let the rule see to that."

Diagnostic Methods: Developmental History.—Developmental history is as useful a method for educational diagnosis as for medical or mechanical or any other form of diagnosis. Go to a doctor with an obscure physical defect and he will enquire about your total past. An automobile repairman asks you to relate just what you did to the car to put it out of order. Take a mentally defective child to a psychologist and he will comb the child's history to see if something in that past may not suggest a diagnosis. The developmental history not only goes back to the pre-natal environment of the child, but to the life of the parents and grandparents. Many fundamental educational defects are not of recent origin. They have been cumulative. They have remained unnoticed for years. Their roots reach far back into the past. A successful diagnosis requires that these roots be traced back to their origins.

Diagnostic Methods: Contrast of Opposites.— Frequently a teacher does not succeed at a diagnosis simply because she does not know what are the customary causes of defects in the ability in question. Suppose, for example, that a pupil is not making satisfactory progress because his method of work is inefficient. A teacher who does not know what methods are and are not efficient is not likely to succeed with this diagnosis.

[11] The clever boy of the class.

A diagnostic method which will help inexperienced teachers is to contrast opposites. The contrast may be between the best and poorest of the class, of pupils in one grade with pupils in a lower or higher grade. This method is to observe the two or three most successful pupils at their work and immediately after to observe the two or three most unsuccessful pupils, or to have both groups trace the process orally, or to test both groups and analyze the results, or to use any other of the diagnostic methods upon both groups at the same time. Diagnosis by contrasting opposites will throw in relief the differences between competent and incompetent pupils and will thus facilitate diagnosis.

Diagnostic Methods: Complete Analysis of Ability.—A complete and thorough analysis of the sensory, mental, and motor processes involved in a given ability is the last resort of the diagnostician. It is the last resort because it is time consuming, and because if it fails the diagnostician can do nothing further. A complete analysis usually requires the combined use of all the previously described diagnostic methods. It utilizes data gleaned from the child's introspections, from observations of his normal work, from the child's oral tracing of the process, from analyses of test results, and from a developmental history.

The technique of diagnosis has been illustrated for arithmetic and reading. The last illustration is for spelling. An excellent illustration for a complete analysis of a school ability is that of Hollingworth and Winford.[12]

"THE PSYCHOLOGICAL EXAMINATION OF POOR SPELLERS

By Leta S. Hollingworth, Assistant Professor of Education, Teachers College

"It is virtually impossible for an educated adult, whose spelling habits have long ago become automatic, to recon-

[12] Leta S. Hollingworth and C. A. Winford, "The Psychology of Special Disability in Spelling"; *Teachers College Record*, March, 1919.

struct from introspection the long, difficult, and complex processes through which he passed in learning to communicate by means of correctly spelled words. Such an adult may gain some idea of what is involved in the spelling process by confronting himself with the task of learning to spell and write words upside down and backwards, but even so the experience of the child is far from duplicated.

"In casting about for material from which to elaborate the analysis of spelling, upon which the psychological examination of poor spellers must be based, it is found that two main sources of information are available. In the first place, we may make controlled observations of children of various ages, who are actually engaged in forming the bonds involved in spelling. In the second place, we may observe experimentally the behavior of those neurological cases, which are characterized by selective loss or enfeeblement of bonds once well established.

"Such observations teach us that the aspect of linguistic attainment, which we call *spelling*, is by no means a simple process, consisting merely in the functioning of a single bond or kind of bonds between a given stimulus and a given response. The process of learning to communicate by means of correctly spelled words ordinarily involves the formation of a series of bonds approximately as follows:

"(1) An object, act, quality, relation, etc., is 'bound' to a certain sound, which has often been repeated while the object is pointed at, the act performed, etc. In order that the bond may become definitely established, it is necessary (*a*) that the individual should be able to identify in consciousness the object, act, quality, etc., and (*b*) that he should be able to recollect the particular vocal sounds which have been associated therewith.

"(2) The sound (word) becomes 'bound' with performance of the very complex muscular act necessary for articulating it.

"(3) Certain printed or written symbols, arbitrarily chosen, visually representing sound combinations, become 'bound' (*a*) with the recognized objects, acts, etc., and (*b*) with their vocal representatives, so that when these symbols are presented to sight, the word can be uttered by the perceiving individual. This is what we should call ability 'to read' the word.

"(4) The separate symbols (letters) become associated with each other in the proper sequence, and have the effect of calling each other up to consciousness in the prescribed order. When this has taken place we say that the individual can *spell orally*.

"(5) The child by a slow, voluntary process 'binds' the visual perception of the separate letters with the muscular movements of arm, hand, and fingers necessary to *copy* the word.

"(6) The child 'binds' the representatives in consciousness of the visual symbols with the motor responses necessary to produce the written word spontaneously, at pleasure.

"This analysis may not be exhaustive, but it provides a foundation on which to construct a scheme for the psychological examination of poor spellers. Obviously, poor spelling may be due to one or another of quite different defects, or to a combination of several defects. In an ability so complex as this there is opportunity for the occurrence of a great variety of deficiencies. In any particular case the underlying cause can be discovered only by means of a psychological examination covering the various mental processes involved. The following outline is based on experimental teaching done at Teachers College during the academic year 1916-1917.[13]

"1. Poor spelling may be due to *sensory defects,* either of the ear or of the eye. If sounds are indistinct, or visual stimuli are vague or distorted, the bonds involving these sensations will be difficult to form. Thus tests of auditory and visual acuity must be given. If any sensory defect is revealed, it should be corrected, if it is corrigible. The necessary tests are described by authors of clinical manuals.

"2. The quality of *general intelligence* must be determined. Failure to spell may be simply one manifestation of general intellectual weakness. For this purpose one of the general intelligence scales is to be used. The Stanford Revision of the Binet-Simon Scale is used by the present writer.

"When we have excluded sensory defects and general intellectual deficiency from the picture, there remains the following possible causes of difficulty:

"3. The bonds which are described in our analysis under (2) may be inadequately or incorrectly developed. This would be *faulty pronunciation.* This is undoubtedly a very prolific cause of poor spelling. Such errors as 'a-f-t-e-r-w-o-o-d-s' for 'afterwards,' 'w-h-e-n-t' for 'went,' 'p-r-e-h-a-p-s' for 'perhaps,' and 'f-a-r-t-h-e-r' for 'father,' will serve to illustrate this point. In our observations on poor

[13] L. S. Hollingworth and C. A. Winford, "The Psychology of Special Disability in Spelling"; *Teachers College Contributions to Education,* No. 88, 1918.

spellers we found such errors by the score, and discovered that the words were pronounced as spelled. Thus the poor speller should be tested for the *pronunciation* of the words which he misspells. It may be that drill in correct pronunciation is what is needed in order to improve his spelling.

"Faulty pronunciation may itself be due to various causes. In the majority of cases it doubtless arises from *false auditory perception,* as in such misspellings as 'hares breath' for 'hair's breadth,' or 'Mail Brothers' for 'Mayo Brothers.' In other cases it arises from *inability to articulate properly,* as with children who stammer or lisp, or have nasal obstructions.

"4. It may be that the weakness lies in the formation of bonds, which we have noted in our analysis under (3). The formation of these bonds involves *visual perception,* which we found to be of first-rate importance in spelling. It has been known for some time that in reading, perceptual factors play a chief rôle. We discovered that among poor spellers error is not distributed at random, but follows certain laws. For instance, there is a constant tendency to shorten words slightly in misspelling them; the influence of any letter over error varies greatly with the position of the letter in the word; the first half of a word has a very great advantage over the last half. From these and other facts it is apparent that weaknesses in visual perception contribute to the failures of many poor spellers. In order to determine whether such is the case with any particular child, it will be necessary to make an analysis of his work, to see whether his errors reveal perceptual weaknesses. If a child can spell the first halves of words correctly, but does not spell the last halves correctly, or if he learns to spell the tops of words correctly, but cannot spell the bottoms of them, the remedy is to bring about readjustments of attention, whereby he will *look* at those portions of words, which formerly he failed, unconsciously, to see.

"5. Poor spelling may be due to sheer *failure to remember—failure to retain* impressions which were originally clearly and correctly perceived. This may mean simply that the child requires an unusually large number of repetitions before he can form the bonds described under (4) in our

analysis; or it may be that his memory span is abnormally short and that he cannot easily associate more than three or four elements together as a unitary sequence. Tests of memory span for various kinds of materials should be instituted in order to gain light on this point. If it appears that his performance is decidedly below the normal for his age, especially when the material is letters, it may be concluded that too brief memory span is probably playing a part in his difficulties. This could be checked up further by an analysis of his spelling, to see to what extent he spells short words correctly, but misspells longer words. In cases where the memory span is brief, emphasis upon syllabication, prefixes, suffixes, and other short units should be helpful. The child might be able to remember three syllables of three letters each, but might be totally unable to retain one word of nine letters. Psychologically these two tasks are very different indeed.

"6. Smedley suggested years ago that there might be a 'rational element' in spelling, whereby *knowledge of the meaning* of words would contribute to the correct spelling of them, in and of itself. Bonds involving meaning are considered in our analysis under (1). In our experimental work we found that children produce many more misspellings in writing words of the meaning of which they are ignorant or uncertain, than they produce in writing words the meaning of which they know. Hence it is of interest to test the child for knowledge of the meaning of words which he misspells. It is necessary to find out whether the words which trouble him are in his vocabulary. It may be that the misspellings which he produces are without content to him. Surely it is conceivable that the absence of a concept might detract from success in arranging the garment in which it should be clad.

"7. *Motor awkwardness and incoördination* may contribute to poor spelling. Here are involved the bonds discussed by us under (5) and (6). In written spelling (with which education is chiefly concerned, since there is but little use for oral spelling in practical life), it is necessary not only to know what symbols are required, but to execute them successfully, with arm, hand, and fingers. Here we must

have recourse to motor tests for steadiness, coördination, and speed of voluntary movement. Occasionally one finds a child who does much better at oral spelling than he does at written spelling. In such cases improvement in handwriting is what is needed, either in rate or in quality. A slow writer may misspell many words if he attempts to hurry.

"8. In the course of our observation we perceived that many of the mistakes of poor spellers are simply *lapses*. These are errors committed by children who 'know better,' who can correct the mistake spontaneously as soon as attention is called to it. There are wide individual differences in the liability to lapse. It is difficult to see what remedial measures may be taken to improve those whose disability is due largely to lapsing, since lapses are not only involuntary, but for the most part unconscious; there is no awareness of them until their primary memory has been lost.

"One might suggest tentatively that children who show this tendency in marked degree should be trained to lay aside for a few minutes all written communications; then to take up their work and look carefully at each word in order to correct all lapses. It is not known experimentally how long an interval must elapse in order that writing may 'get cold,' so that lapses may be detected by the author of them. A few minutes will probably suffice.

"9. *Transfer of habits previously acquired* is occasionally the cause of misspelling. We found, for example, one poor speller, who had previously learned in a phonetic language. He carried over this habit into English spelling, and it was very difficult for him to adjust himself. The possible existence of such an influence is to be determined by taking the child's school history.

"10. Sometimes it happens that the errors of a child are largely of one particular kind. Such *idiosyncrasies* may be exemplified by the case of a child who had a strong tendency to add final 'e' to all words; and by the case of another who was addicted to intrusive consonants, especially 'm' and 'n.' These idiosyncrasies may doubtless be traced to their source in every case by a patient analysis of the mental contents of the child. The cause of error will be different in every case. It is impossible to generalize about idiosyncrasies.

"11. After all of the foregoing factors have been considered, there still remains the possibility that the failure to learn is due wholly or partially to temperamental traits —indifference, carelessness, lack of motivation, distaste for intellectual drudgery. English spelling calls largely for rote learning. It can be acquired only by the formation of thousands of specific bonds, arbitrarily prescribed. Its pursuit is extremely tedious at best. Thus many children will be temperamentally ill adapted to become good spellers.

"Disability in spelling may result from any one of the defects which we have outlined, or from any combination of two or more of them. It is apparent that the psychological examination of a poor speller is neither a brief nor a simple task. The direct examination of the individual should furthermore be supplemented by a family history, a developmental history, and a school history. In some cases special defect in spelling appears to be hereditary. Stephenson,[14] for instance, has reported six cases of inability to read and spell, which occurred in three generations of one family. In some of our own cases relatives have been affected with linguistic disabilities.

"A developmental history will reveal whether the child was backward in speech, whether he has or has had any speech defects, and whether he has been affected by any illness that might conceivably have produced localized leisons in the central nervous system, or have affected linguistic ability in any other way. One of our cases, for example, had suffered a paralysis of the soft palate following diphtheria, which had for some time interfered with articulation. Another had a history of having been tongue-tied till he was eight years old. Such facts may be of considerable interest in a given case.

"A school history is essential in order to determine whether progress in school subjects other than spelling has been normal, whether learning in a language other than English has taken place, and whether the disability has operated to cause general retardation in school status.

"The question naturally arises as to whether the difficulty is always remediable when located. It is quite possible that

[14] S. Stephenson, "Six Cases of Congenital Word-Blindness Affecting Three Generations of One Family"; *Opthalmascope*, August, 1907.

there exist cases where the necessary bonds cannot all be formed, even with the maximum of practice and effort. Experimental teaching has not yet been undertaken to an extent which would give the answer to the question. In those rare cases where the disability is very extreme, in a child of good general capacity, it is probably wise to make some special provision for oral recitation and examination and thus to allow the child to make progress in school, rather than to keep him back year after year on account of his disability."

Prerequisites of Skill in Diagnosis.—Success as a diagnostician requires: (1) A knowledge of the usual causes of usual defects in the various abilities developed by the school. (2) Eyes to see and training or experience to interpret subtle behavior as evidence of the operation of known causes. (3) A technique which will bring otherwise invisible hints to the surface. (4) A knowledge of what remedial measures to prescribe for a given diagnosis.

Summarized below are certain basic causes which are responsible for many defects and whose operation is not confined to any one subject. Just as pestilences can usually be traced back to a few sources, so many diagnostic traits, irrespective of the abilities from which they start, lead back to a few basic causes, especially when the defect being diagnosed is an annoyingly persistent one. Before anyone attempts diagnosis he should have a knowledge of the more common fundamental breeders of ability defects.

Insufficient practice.—In some pupils a given ability does not function at all, simply because they have never studied to develop the ability, or it functions imperfectly because they have not had enough study and practice. This condition need cause no special concern for it is easily remedied. The time for real concern comes when a normal amount of study and practice fails to eliminate the absence or imperfection of functioning.

Improper methods of work.—There may be an optimum method of work. Pupils who differ in type or temperament

may require different methods or again there may be an optimum method for all pupils. At any rate many pupils are working below par because they are employing ineffective methods.

A special case of improper methods of work occurs in those abilities where speed and quality are intimately related. It may be that a pupil's ability is functioning imperfectly because it is functioning either too speedily or not speedily enough.

Deficiency in fundamental skills.—Deficiency may mean either absence of sufficient skill or absence of sufficient transfer of skill to the new situation or both. The mental processes cannot flower into appreciation of literature, nor is the mind free to reflect upon the great principles of history, geography, science, mathematical problems and other higher stages in education until the underlying skills are both made automatic and transferable. The youth who does not come from a cultured home and whose learning has been hastily grafted on an ignorant home training, is barely conscious of his own ideas when addressing a cultured audience and scarcely enjoys what he eats when dining with a cultured family. All his attention is concentrated upon watching lest he "gabble like a goose," or upon observing lest he use the wrong spoon. The pupil who stumbles in his reading halts in his history. The remedy is to make the basic skills automatic.

Absence of interest.—The importance of interest or purpose in developing ability cannot easily be over-emphasized. There are more failures due to failure of interest than this world dreams of.

Physical defects.—The diagnosis of any ability should carefully consider physical factors. In the case of many pupils food for their minds will not facilitate their school progress nearly so much as food for their stomachs. No diagnosis should omit a careful examination of sense organs, particularly the eyes and ears. Just as "rivers of mercy do not flow into the world through rye-straws," so we do

not have an educational flood when knowledge and experience must trickle through choked sense organs. Instruction cannot possibly be more than 50 per cent efficient when the child hears only 50 per cent of what is said to him and sees only 50 per cent of what he looks at.

Again, diagnosis should consider the condition of the pupil's response mechanism. What goes in through the sense organs must come out through the response organs before the educative cycle is complete. More improvement in molding, drawing, painting, writing, manual arts and sports might conceivably be secured through correction of defects of muscular coördination than through direct instruction in the abilities in question.

> "Thy body at its best
> How far can that project thy soul on its lone way."

Subnormal intelligence.—Low native intelligence is the preëminent cause of ability defects. Intelligence is the very tap-root of Igdrasil. Just as injury to the tiny pituitary body causes stunted stature, marked adiposity, imperfect sexual development and other profound changes, so a defective intelligence casts its blight upon many or all abilities. Because of its ubiquity and its probable unimprovability, this cause of defects has special significance. Its importance is not always understood by the superficial diagnostician, because the superficial diagnostician does not carry the process of diagnosis far enough. Unsatisfactory work in history may be traced to imperfect reading ability. But why is the reading ability imperfect? In many instances it will be found that reading ability is imperfect because of low native intelligence. Whenever retardation is general, and whenever there is relative unimprovability, it is well to test for intelligence.

CHAPTER IV

MEASUREMENT IN TEACHING

I. The Use of Practice Tests

Practice Tests Described.—After the initial tests are given the teacher should teach reading according to the best pedagogical procedure. This does not mean that measurement should be eliminated at this stage, for teaching and testing should always be continuously intermingled. In the case of the fundamental skills it is advisable to teach almost entirely by testing. Practice tests in arithmetic, handwriting, etc., make this possible. At the time of writing, practice tests have not been constructed in the field of reading. Until such tests are constructed it would be advantageous to have pupils read material cleverly selected so pupils will manifest in some overt fashion the extent to which they comprehend the material being read. Effective teaching requires this constant measurement of comprehension.

A knowledge of the fundamental psychology of learning is necessary but not always sufficient for effective teaching. It is frequently difficult to translate principle into practice. A detailed translation would require several volumes. Only enough space can be spared to describe three important contributions of educational measurement toward bridging this gap between principles and practice, namely, *practice tests, informal tests, and standardized scales.*

Though practice tests are in use in both elementary and high schools [1] it will suffice to describe a typical elementary-school practice test—Courtis Standard Practice Tests in

[1] I. M. Allen, "Experiments in Supervised Study," *School Review*, June, 1919.

Arithmetic.[2] A set of these practice tests consists of 48 stiff cards which make 48 lessons. Each lesson, except lessons 13, 30, 31 and 44 which are test cards, and lessons 45, 46, 47 and 48 which are study cards, contains just one type of example. The lessons begin with simple examples and gradually become more complex, each additional lesson representing just one additional difficulty. When the pupil has mastered the forty lessons, he has mastered all the difficulties in the addition, subtraction, multiplication, and division of whole numbers. There is one set of practice lessons for each pupil.

Along with the practice lessons comes a Student's Practice Pad for each pupil. The practice pad contains sheets of tissue paper. The pupil inserts a lesson card into the pad and under a sheet of tissue paper. This permits the pupil to see the example and at the same time do all work on the tissue paper, thus enabling the lesson card to be used from year to year. The student's practice pad also contains sheets upon which a pupil can keep a daily tabular and graphic record of achievement and progress.

Along with both practice lessons and practice pad comes a Teacher's Manual, which gives detailed instructions for the proper use of practice lessons and practice pads and warnings against their improper use by over-zealous teachers. The manual also gives much helpful advice about how to diagnose and remedy pupil defects in the four fundamental processes. The manual also contains record sheets which enable the teacher to keep a continuous record of each pupil's work.

The essential steps in the procedure of using these practice tests follow:

1. All pupils are given test card 13 which contains all the difficulties found in lessons 1 to 13. Each pupil slips the test card, examples up, under the topmost sheet of tissue paper in his practice pad. At the signal all begin work and continue until the signal is given to stop.

[2] Sold by World Book Company, Yonkers, N. Y.

2. Pupils exchange papers and score each other as the teacher calls the correct answers.

3. All pupils who make satisfactory scores are excused from lessons 1 to 13. Sometimes the test is given twice to make results reliable. Sometimes the excused pupils may do something else until the backward pupils catch up or they may take the next test and the next until a point is reached where they need to study.

4. All pupils not excused from drill take lesson 1. If they make a satisfactory score on lesson 1, the next day they take lesson, 2, and so on.

5. Those who fail on lesson 1 continue studying it and taking it until a satisfactory score is made.

6. As soon as a pupil finishes lesson 12 he takes test 13 again as final proof of his mastery of the preceding lessons. He may work on something else until the others catch up or he may proceed.

7. As soon as about 90 per cent of the class, including those who originally passed, have finished test 13, they take test 30. Those who pass test 30 are excused, and those who do not, drill upon lessons 14 to 30 as described above.

8. The teacher keeps a daily record of what each pupil achieves, watches to see that there is no cheating, makes diagnoses and applies remedies where they are needed and *only where they are needed,* stimulates good work on the part of all, sees that pupils keep their own records in good condition, and occasionally rescores the pupils' papers in order to keep their standard of scoring high.

All the regular lesson cards have answers on the back, hence pupils may score themselves or each other by simply turning the lesson card over and reinserting it under the tissue paper. The teacher's attention is thus freed for the real work of individual instruction, since no papers are handed in to her except those which the pupil himself judges to be perfect.

Practice Tests Individualize Instruction.– Mass instruction is highly inefficient, and this is particularly the

case with skills. The interests of study, instruction and supervision are identical. All focus upon study. Study is highly individual. Instruction must be equally individual if it is to be efficient. Mass instruction aims at everybody. It frequently hits nobody.

The amœba has three types of reactions produced by three types of stimuli. There are, first, positive stimuli in the form of satisfying food and the like. The amœba reacts by advancing toward these stimuli. The teacher uses positive stimuli to attract pupils toward good habits of work. There is, second, negative stimuli to which the amœba reacts by retreating. The teacher uses negative stimuli to drive the pupil out of bad habits of work. There is, third, neutral stimuli which produce neutral reactions in the amœba, for neutral stimuli do not stimulate at all and neutral reactions simply mean no reactions at all. It is the teacher's ambition to become so efficient that every word she speaks or move she makes will be a positive or negative stimulus depending upon her choice. But in mass instruction most of the stimuli are neutral stimuli. Our professor of literature was right when he said that teaching the class was "like trying to pour water from a gallon bucket into small-necked bottles." Most of his stimuli were neutral, partly because of lack of capacity on our part, partly because he was employing stimuli which were neutral to most, negative to some, and positive to only a few. Individual differences are so great that wherever possible mass instruction should give way to individual instruction. Practice tests are a device for individualizing instruction. Without the aid of some such device individual instruction is impracticable.

Practice tests automatically adapt the work to the ability of each pupil and thus enable each pupil to begin at that point which means neither reteaching nor premature teaching. This is accomplished by means of the initial inventory tests. Test 13 serves this function in the case of the Courtis Practice Tests.

Practice tests permit each pupil to work according to his own methods and help him to find his best method. It is surprising how varied are the methods by which pupils learn such narrow functions as addition, subtraction, multiplication, and division. Kirby [3] has shown not only that what is the best method for one pupil is not always the best method for another, but also that pupils frequently do not discover their best method and best rate of work until they are under the pressure of raising their score.

Finally, practice tests permit each pupil to advance at his own rate. Every study of the varied rates of progress for pupils in the same class has revealed the need of some teaching method which makes provision for individual differences in this respect.

Practice Tests Strengthen the Purpose to Improve.—Practice tests motivate the learning process by making visible both distant and immediate goals and by providing a method whereby a pupil can measure his rate of progress toward these goals. Every pupil keeps a record of each day's achievement and draws a graph showing his progress. These provisions motivate through their appeal to basic instincts. The instinct of rivalry is so strong that work is turned into play by the simple process of introducing into it this element of rivalry. Practice tests not only make possible a rivalry between individuals, which is probably the world's most ubiquitous form of motivation, but they also make possible higher types of rivalry, namely, rivalry with one's own past record, and the rivalry of one group with another.

This provision of practice tests for the keeping of scores is prerequisite both to intense effort and real happiness in school work. The games at which both children and adults work hardest and are happiest are invariably games where a score is kept. Generally speaking conventional education

[3] Thomas J. Kirby, *Practice in the Case of School Children;* Teachers College, Columbia University, New York, 1913.

does not keep scores. A sort of score is occasionally reported, but these scores are purely relative. They do not show how much each individual has surpassed his previous record. They show which pupil is relatively best and so on to poorest. What stimulus is that to pupils who know they cannot hope to outstrip a more capable competitor? And what stimulus is it to the victor who knows that victory comes without much exertion due to his native superiority?

Practice tests motivate learning by throwing responsibility for promotion, or the attainment of the goal, upon the pupil. Every idle minute puts off the day when the goal will be reached and every industrious moment hastens the coming of the day, and what is important, the pupil is made to clearly perceive this intimate relation and is forced to recognize the fairness and justice of it. Just as certain as a pupil idles he will be punished and just as sure as he works he will be rewarded.

Practice Tests Secure a Maximum of Exercise.—The second fundamental law of learning is, according to Thorndike, the law of exercise. When purpose is strong or when the law of effect is appropriately utilized and when exercise is abundant we have the optimum conditions for rapid progress. Here is the way I once taught addition to a class of forty pupils.

"Will each pupil copy on a sheet of paper the addition examples which I shall read to you, five examples to the row?" And then,

"Mary, you give orally the answers to the examples in the first row." And then,

"John, you take the second row," etc.

Each patiently or mischievously, according to his nature, waited until his turn came to begin. Only one pupil's neurons were exercising at a time, because I told each one just exactly where the preceding one stopped. Subsequent observations of other teachers have shown that my stupidity was not an isolated case. This one-out-of-forty

sort of exercise is quite common. Had I used modern practice tests, probably without knowing it, I would have multiplied my efficiency just forty times.

Practice Tests Facilitate Aid and Diagnosis.—Practice tests bring swift aid to the pupil who needs it, and prevent teaching when it is not needed. Effort expended which brings no return in terms of progress brings discouragement. When discouragement reaches a certain stage effort ceases. Under ordinary conditions pupils sometimes remain for years undiscovered in the Slough of Despond. When the pupil's curve of progress ceases to rise to reward his effort, a teacher is needed. For the teacher to help at any other time would probably be to waste her time and injure the pupil. When to teach is instantly revealed by the curve of progress graphed by the pupil.

Practice tests facilitate diagnosis. Successful diagnosis requires the teacher to discover the exact location of the difficulty and the exact cause of the difficulty. Like tracer bullets, the pupil's daily scores leave behind a fiery trail which instantly reveals the location of the difficulty. The very following of this trail helps to eliminate probable explanations and thus facilitates diagnosis.

The chief danger from practice tests is not that they will cause too much emphasis upon drill, because the accompanying manuals allot a conservative time and constantly urge teachers not to exceed this time. The chief danger is that teachers will consider practice tests as something apart, so that the abilities developed by them will not function in life situations. The use of practice tests should grow out of genuine situations and should be continually associated with genuine situations. There comes a time in the execution of projects where the pupil realizes that his skill is inadequate. It is the function of practice tests to repair this inadequacy in the most economical and interesting way.

II. The Use of Informal Examinations [4]

Importance of Examinations.—There are in the United States about 700,000 elementary school, high school, and college teachers. It is a conservative estimate that each teacher gives on the average twenty examinations a year. This makes 14,000,000 examinations each year. The time required to construct, give, and score each examination will average, say, three hours. This means that annually about 42,000,000 hours are spent examining pupils. Even though our estimate is doubly generous, the hours would still be sufficient to show the enormous importance of examinations. Without a doubt, examinations are and will be for some time and may possibly always remain the most important form of educational measurement. Since this is so, it may seem that those of us who are interested in educational measurement have, in our enthusiasm for constructing and standardizing tests, neglected the traditional type of educational measurement. Really, however, this has not been neglect on our part, for standard tests are nothing but improved examinations. Furthermore, we have been learning new techniques which will in time react to improve the making of examinations. The purpose of the next few pages is to show teachers how they may make use of one of these new techniques of scientific testing not only to improve certain kinds of examinations, but also to make examinations a real pleasure instead of an onerous task to both teacher and pupils.

Sample True-False Examination.—The scattered examination shown below is designed to test a pupil's knowledge of certain facts concerning the physical features of the United States. In actual practice a teacher will usually test on a much narrower topic. We have purposely written this examination hastily in order that it might illustrate certain crudities of construction. Any teacher in the elementary

[4] A modified form of this section first appeared thus: Wm. A. McCall, "A New Kind of School Examination"; *Journal of Educational Research*, January, 1920.

school could do as well and most teachers could do better. The same technique is equally useful to high school and college teachers.

The examination as presented here assumes that the statements whose truth and falsity are to be determined by the pupils have been mimeographed so that a copy of the examination could be placed in the hands of each pupil. The sample examination given below has been worked through by a pupil and been scored by a pupil or the teacher. The underlining was done by a pupil. The check, cross and zero mean respectively that the pupil's answer is correct, incorrect or omitted. Only enough of the examination is shown below to illustrate the procedure.

SAMPLE EXAMINATION ON UNITED STATES

Some of the following twenty statements are true and some are false. When the statement is true draw a line under True; when it is false draw a line under False. Be sure to make a mark for every statement. If you do not know, guess.

1.	In general the mountain ranges run east and west.	True	False	✓
2.	Most of the rivers flow north.	True	False	✓
3.	Mt. Mitchel is the highest point east of the Mississippi River.	True	False	X
4.	Mt. Washington is higher than Mt. Mitchel.	True	False	X
5.	The Catskill Mountains are in Maine.	True	False	✓
6.	The Cascade Mountains are nearer the Pacific Ocean than the Rocky Mountains.	True	False	X
7.	The Rocky Mountains are nearer the Pacific Ocean than the Appalachian Mountains.	True	False	✓
8.	The Blue Ridge is in the Rocky Mountains.	True	False	✓
9.	There are more active volcanoes in the west than in the east.	True	False	✓
10.	"Old Faithful" is the name of a cyclone which sweeps upward from Texas into Oklahoma.	True	False	X
11.	The "Grand Canyon" was cut through the Cumberland Plateau by the Susquehanna River.	True	False	✓
12.	Pike's Peak is in the Rocky Mountains.	True	False	✓
13.	The Mississippi River flows into the Great Lakes.	True	False	✓
14.	All the following are tributaries of the Mississippi River: Arkansas, Missouri, Ohio.	True	False	✓
15.	The Big Sandy is the biggest river in the United States.	True	False	X

16. The Atlantic Ocean is to the east and the Pacific Ocean to the west.	True	False	O
17. Canada is to the south and the Gulf of Mexico to the north.	True	False	√
18. The great lakes are five in number.	True	False	√
19. It is easier to sink while swimming in the largest lake east than in the largest west of the Mississippi.	True	False	√
20. The central portion of the United States is on the whole more level than the eastern or western portion.	True	False	√

How to Compute Score for True-False Examination.—

Number of correct underlinings = 14
Number of incorrect underlinings = 5
Number of omissions = 1

(A) Pupil's score = number correct — number wrong.
Pupil's scores = 14 — 5 = 9

Let us consider first the reason for expressing a pupil's score as the number correct minus the number wrong. Imagine a pupil who is absolutely innocent of any knowledge of the physical features of the United States. Were such a pupil to take the above test and were he to mark every statement he would according to the theory of chance mark ten statements correctly and ten incorrectly. The chances of his guessing right or wrong are fifty-fifty or one to one. His score on the above test would be:

$$\text{Score} = 10 - 10 = 0$$

In short, the pupil's knowledge is zero and the method of computing his score gives him zero. Suppose instead that he knows ten statements and guesses at the other ten. Of the ten guessed at he would, according to chance, get five correct and five wrong. That is, even though his real knowledge is ten he will show fifteen correct (10 + 5) and five incorrect. The method of computing his score brings out his real knowledge.

$$\text{Score} = 15 - 5 = 10$$

A pupil who marks every statement correctly makes a perfect score, viz.:

$$\text{Score} = 20 - 0 = 20$$

Observe that no account is taken of omissions. Only the corrects and incorrects figure in the pupil's score. When the time allowed the pupils to take the test is made short in order to test each pupil's speed of work there will, of course, be many papers showing several omissions each. In all such cases omissions should be ignored, just as we have done above, in computing scores. Even when the time allowed for the test is ample for each pupil to mark every statement, there will still be an occasional instance of omission due to carelessness or misunderstanding of instructions or a puritanic conscience against increasing the score by gamble guess-work even when the instructions urge guessing.

When the time is ample for even the slowest pupils and when all are instructed to mark every statement, it is much more convenient to compute a pupil's score according to the following formula:

(B) Score = (number of statements) — 2 (number marked incorrectly)

If there are twenty statements in the test and if five are marked incorrectly,

$$\text{Score} = (20) - 2\,(5) = 10$$

Formula (A) gives the same results

$$\text{Score} = 15 - 5 = 10$$

That both formulæ give identical results provided there are no omissions may be shown viz.:

Let T = Total number of statements, R = number right, and W = number wrong.

Then

$$\text{Score} = R - W \qquad \text{(Formula A)}$$

$$R + W = T$$

$$R = T - W$$

Substituting in Formula A

Score = T — W — W = T — 2 W (Formula B)

Formula (A) is basic and should be used when there are omissions. Formula (B) should be preferred when there are no omissions or when they are present only in negligible amount. Formula (B) is much more convenient. The first number is always the same and since the second number is the total statements marked incorrectly, it is only necessary to score and total the errors.

It is very difficult for some people to believe that such a test as has been outlined above does anything more than give the highest score to the luckiest guesser. They look with the eye of suspicion upon this thing we call *chance.* I once tossed pennies for heads or tails 50,000 times. The results came out 25,000 heads and 24,999 tails. Had there not been a miscount somewhere the two would doubtless have come out exactly even. I had occasion to *watch* two summer-school teachers in that nerve-racking game of chance called *matching pennies.* They matched for several minutes daily. The last heard they were still matching pennies and chance had prevented either from getting complete possession of the other's 100 pennies. Chance is fatally exact when the pennies or the statements in the test are numerous. The opportunities for injustice in score multiply in proportion as the number of statements is reduced. Hence there should be as many statements in the test as practical limitations will permit.

Detailed Construction of True-False Examination.— There are a few suggestions which will help teachers in constructing the *True-False* test. In the first place the teacher should so construct the test that it will contain approximately the same number of true and false statements. A clever pupil may get a higher score than he deserves if he discovers there are many more true statements than false statements in the test or vice versa. Suppose there are many more true statements than false statements and suppose some pupil discovers this by observing the statements that he knows, or by observing the teacher's bias for writing

true statements instead of false ones. Naturally when he does not know what to mark he will mark *True,* thereby securing a larger score than his ability justifies. Probably it is by just such utilization of the errors of others that the intelligent get through life so much more smoothly than the stupid. On the other hand, the teacher should not have exactly the same number of true and false statements each time, because this will invite clever pupils to count back to see how many more true statements have been marked than false statements. Sometimes there should be more true statements, sometimes more false statements, sometimes the same number of each. Any regularity of plan should be carefully avoided. An English admiral complimented the skill of German submarine commanders by saying they were masters of irregularity. All the true statements should not come first, neither should the true and false statements be alternated as a regular plan. Let chance determine what shall be true and what shall be false and in what order the true and the false shall come.

Second, the teacher should be careful to keep out of the test all ambiguous statements. Statement number 18 in the sample test is somewhat ambiguous. It says: "The great lakes are five in number." Since *great lakes* is not capitalized a pupil might very legitimately interpret this to include the Great Salt Lake and others. It will later be difficult to satisfy this pupil that his score should suffer because of the construction he gave this sentence. If the teacher will study her mistakes in this respect she will soon learn how to reduce such ambiguities. As any teacher can testify, the danger of ambiguities of wording are not peculiar to this test. This type of test does not, however, give a pupil an opportunity to reveal just what interpretation he places upon each statement unless the teacher follows the procedure of having pupils score their own or each other's paper. Self-scoring will reveal all cases of ambiguity. Statements which are particularly flagrant in this respect can be omitted in scoring.

Third, the teacher should inspect not only this but any sort of test from the point of view of just what the test measures. Statement 19 in the sample test illustrates this point. The purpose is to test whether the pupil knows that the largest lake west of the Mississippi River contains more salt than the largest lake east of the Mississippi. Instead of measuring this I may be testing whether a pupil knows that it is easier to sink in fresh water than in salt water. Complex wording, unfamiliar terms, the use of negatives, all tend to make the test a linguistic one. Simple, brief statements without negatives are best.

Fourth, the teacher may so construct the examination as to force pupils to guess wrong due to the power of suggestion. This probably explains why statement 15 was marked wrongly. The pupil doubtless argued to himself that since the river is named the Big Sandy it probably is the biggest river in the United States. The influence of having many suggestive statements in the test is to make the examination more difficult. It operates to give to the pupil who knows nothing at all in the test a large negative score instead of a zero score and it penalizes rather heavily the pupil who does much guessing, for every time he allows himself to be suggested in the wrong direction a point is subtracted from the score he has already made by what knowledge he has. In other words, the suggestive statements make the gap between those who know much and those who know little wider than it otherwise would be. Whether a pupil should be specially penalized for yielding to suggestion is an arguable question. There may be situations where it is eminently desirable to determine whether pupils know what they know so well as to be able to resist suggestion. In general, however, it is best to avoid suggestive statements. The ideal should be to construct the examination so that any pupil who knows absolutely nothing about the test will make a score of zero.

In sum, the examination should harmonize with the following suggestions:

1. Have approximately the same number of true and false statements and have them arranged in chance order.
2. Avoid ambiguous statements.
3. Avoid suggestive statements.
4. Avoid trivial statements lest they induce wrong habits of study.
5. Avoid the use of negatives.
6. Make the statements brief.
7. See that one statement does not answer a preceding one.

Methods of Applying True-False Test.—So much for the construction of the examination. How shall it be applied? The best way is to print, mimeograph, or otherwise duplicate, the examination, and place a copy in the hands of each pupil. But there are numerous schools which lack duplicating machines. For teachers in these schools some other means of applying the test must be found. Any one of the following methods may be used. First, the entire test may be copied word for word by the pupils and then marked. This is tedious and time-consuming. Second, the entire test may be written on the blackboard by the teacher. Each pupil could number a blank page of paper to correspond to the numbered statements, and then write *True* or *False* after the appropriate numbers. The only objection to this suggestion is the inconvenience of writing all the statements on the blackboard. Third, the pupils may be asked to copy on blank paper, 1, 2, 3 and so on, according to the number of statements. The teacher can then read orally statement 1 and instruct the pupils to make a check after the number 1 on their paper if the statement is true, but to make a cross if the statement is false. This is easily the most convenient way to give the examination. The chief objection to this final method is the difficulty some pupils have in apprehending statements presented orally, particularly if they are long and complicated. When the statement is presented visually the pupil has an opportunity to go back to it enough times to exhaust his possibility of

understanding it. By one or another of these methods it is possible for any teacher anywhere to make use of this type of examination.

Scoring of True-False Examination.—How shall the True-False examinations be scored? If a copy of the test has been placed in the hands of each pupil, the teacher can take an unused test sheet, fill it out correctly, lay the correct column of answers beside each pupil's column of answers, and quickly mark whether the pupil's answers are correct or incorrect. If a copy of the test has not been placed in the hands of each pupil, but each has instead written *True* or *False,* or made a check or cross after the number of each statement, the teacher can take a page of paper similar to that on which each pupil has indicated his answers, copy the numbers just as they are and just as they are spaced on the pupils' papers, write after each number the correct answers to the statement of that number, place this column of correct answers beside the column of pupil answers and mark those which are correct and incorrect. This last scoring method presupposes that pupils have used ruled paper, and that each has written his numbers in a vertical column according to a particular spacing recommended by the teacher. Last and best, each pupil can score his own or his neighbor's paper. It is better for him to score his own.

If the method of pupil scoring is adopted, the teacher should read the correct answers while the pupil checks his own. If the pupil does not have a copy of the statements before him, the teacher should read each statement before giving the correct answers, in order that the pupil may know what statements he got correct or incorrect. When all the pupils' answers have been marked and when all their scores have been computed and recorded on their examination paper, the teacher should ask all the pupils who missed statement number 1 to hold up their hands, and then all pupils who missed number 2 to hold up their hands, and so on. The teacher should make a record of the number of pupils missing each statement, and then collect all papers.

But pupils will cheat. To be sure some will cheat. It will advantage us nothing to delude ourselves into the belief that cheating will not occur. To do so would be to join the peerage of the ostrich that is fallaciously reported to stick its head into the sand and think itself safe, or of the partridge which dives into a snow bank and feels as secure of its safety as the hunter feels of his game. It would advantage us still less to compel honesty by so arranging all educational situations that there is no opportunity to be dishonest. The chances that the world will be so tender of a pupil's weakness are very few indeed. If a pupil has it in him to be dishonest, it is a genuine kindness for the teacher to find it out. The benevolent birch removes less epidermis than the rod of the law.

However determined, the scores for the pupils may be left either in their original form or they may be scaled. A later chapter describes a very simple method of converting crude scores on an examination into scale scores in terms of a common basic unit called T. The few moments consumed in making this transmutation are more than repaid by having records for each child which are comparable from examination to examination.

Advantages of True-False Examination.—But why should this sort of examination be used at all? Wherein is it better than the examination method in common use? In the first place, the *True-False* examination permits a teacher to cover a wider field of subject matter or a wider range of ability per unit of time. It may be made more representative of the total field of the pupils' study and hence be a fairer measure of the pupil. In the case of the traditional examination the teacher is forced to select a very small number of questions. When we were students almost as much of our ingenuity went into divining the kind of examination questions the teacher would ask as in reviewing for examination. Now that we are teachers we have no reason to suppose that this practice has ceased.

The use of this type of examination is likely to improve

the relation between teacher and pupils. The traditional examination endangers a pleasant relationship because pupils more or less justly suspect that the score they make depends almost as much upon their conduct as upon their product. The proposed test convinces a pupil that the score he gets is the score he deserves. Such a conviction is a real event in a pupil's educational career.

The *True-False* examination is more enjoyed by the pupils. The pupils enjoy it more because it offers an opportunity for a contest where the rules are fair, and because it offers them a chance for a large degree of participation in the examination. It is agonizing for a pupil to describe at great length a knowledge which he does not possess in hopes that his command of English will camouflage his lack of information. Here is a question which was asked in a recent examination in educational measurement. "Which three of the tests described by Whipple do you think would be of most service in an elementary school, if your school had a school psychologist to apply them?" Consider the perspiration it must have cost a student to perpetrate this answer:

"The tests described by Whipple embraced most of the difficulties that would be embraced in problems of classroom instruction. I think his tests embrace a great variety of methods of approach and it seems difficult for me to think of just three to whom the presence of a psychologist in a school would give help. I would think it would be the tests in which knowledge of the workings of a child's mind and its growth and development would be most apparent since those not particularly trained might focus on others not of this kind. I fear it would be unwise to specifically mention just three when the number is so great which would fulfill all these requirements. Every teacher to be a psychologist would help all classroom measurement work of whatever kind greatly, I know; since we cannot know of the influence of a test upon any group except by the mental reaction produced."

The *True-False* examination is more enjoyed by the teacher. The scoring is easy, rapid, and automatic when she does the scoring, and far more rapid when the pupils do the scoring. The pupils cannot well assist in scoring the traditional examination, and for the teacher to score forty verbose examination papers is time-consuming drudgery. Every moment of the time while scoring, the teacher must be profoundly concentrating upon what she is reading, for much of the time she must be separating the chaff from the wheat where the chaff is cleverly painted to look like wheat. And along with this is a continual emotional strain caused by her resistance to the temptation to underscore some and overscore others.

The *True-False* examination is more educative for the pupils. The proposition that pupil scoring will relieve the teacher of much obnoxious drudgery, does not justify the inference frequently made that what is non-educative drudgery for the teacher will also be non-educative drudgery for the pupils. On the contrary the most favorable teaching opportunity that ever comes to a teacher is the period immediately following an examination. The pupil's interest to know what parts of the examination he missed and what he got correct is then at white heat. Witness the interested discussion among pupils immediately following an examination. It is inexcusable neglect of an educational opportunity not to capitalize these precious moments for correcting erroneous ideas, clinching right ideas, and filling up mental spaces where ideas are not. These values can best be realized by having each pupil score his own paper and by stopping to discuss points where pupils have trouble. Of course not every correct answer indicates knowledge, but the pupil himself usually knows when he knows. This examination is also more educative, because it is likely to be given more frequently. The experience of Kirby, Courtis and others with practice tests shows that a pupil learns more during testing periods than during teaching periods. We really teach when we test. This examination covering as it can a

wide range is an ideal method of review. It reveals to the pupils just where their difficulties lie. Testing is one of the best ways of teaching.

The *True-False* examination gives the teacher a fuller knowledge of conditions. The educative value of testing is so great that testing should be much more frequent than is now the case. Now that a method of testing is available which involves no drudgery to anyone, testing is likely to become more frequent, and this means more complete and timely information about the abilities and difficulties of the various pupils, and about the successes and failures of teaching efforts. It has already been suggested that the teacher keep a record of the number or per cent of pupils missing each statement in the examination. This record will show what things have been well learned or poorly learned and well taught or poorly taught. Also it is a good thing for a teacher to check her own efficiency in general. This can be done by finding the average of the scores of all the pupils and by comparing this average with the total number of statements in the examination or at least the total number of facts the teacher has really attempted to teach the pupils. If the average score is 20 out of a possible 40, the teacher's efficiency is 50%. Most teachers will be chagrined to find, if they use truly representative statements in their examination, that their efficiency is below 50%. Similarly, a pupil's efficiency may be determined by the per cent of statements he got correct out of the total number of statements the teacher has a right to expect him to get. Before the examination is given the teacher should decide what statements she has a right to expect the pupils to get correct. This same number should then be used for computing both pupil and teacher efficiency.

Finally, the *True-False* examination is a genuine honesty test, and shows the beginnings of a technique for measuring in satisfactory fashion this valuable character trait. Occasional and unannounced rescoring of each pupil's paper by his neighbor will catch the persistent cheat. It is better

that he be discovered in school than in court. His discipline can usually be left to his fellow pupils, over whom he was attempting to gain an advantage dishonestly.

There has been listed what appear to be the chief advantages of this type of examination or any other examination which is similarly objective. Most of these claims rest upon logical probability and a limited experience and not upon experimental data. This last is needed and will follow in time.

There are some limitations which have not been discussed. It is claimed first that this examination does not require the pupil to demonstrate a power to organize his materials. This is true in the sense that the pupil does not *describe in writing* a complicated mental organization but a statement can be so worded as to require an exceedingly complex mental organization before a correct answer can be unfailingly given. Consider the mental organization that must precede a correct answer to this simple statement: "If the trade winds blew east Peru would have luxuriant flora." If it is desired to test a pupil's power to word his thought a composition test may be given.

Again, it is claimed that this examination can test knowledge but not skill, knowledge but not the ability to do. Even skills can be tested by this examination. To reason that trade winds blowing east would be warm, would absorb moisture from the Pacific, would become chilled in passing over the Andes, would consequently deposit a heavy rainfall for Peru, which taken in conjunction with the equatorial climate would produce a luxuriant flora, is one sort of skill which this examination will test. Mathematical skills and the like which are too complicated to describe may be tested in at least two ways, though there are better ways. An example or problem may be stated together with an answer. The pupil's task would be to determine by working the problem whether the answer given is true or false. Or instead, the teacher can work the problem on the blackboard

for all the pupils and have them indicate whether her process was correct or incorrect.

Finally, it is claimed that the teacher needs to know why a pupil is unable to answer the question about Peru and its flora. The *True-False* examination does not show just where the pupil's reasoning process went wrong or stopped altogether. It is not diagnostic. This criticism has some force. An examination should be as diagnostic as possible. If a teacher wished to know where the pupil's process broke down she could give a subsequent, more detailed examination of this type. The statement, "The trade winds are warm winds," or "Warm winds have a larger capacity for water than cool winds," etc., would reveal whether the pupils were acquainted with the basic principles, facts and the like necessary to reason out the correct answer to: "If the trade winds blew east Peru would have a luxuriant flora."

The traditional examination has a certain advantage which will doubtless continue its existence. The *True-False* examination is but a herald of the newer and better types of examination to be. But even now we have in the *True-False* examination one that may be used by any teacher anywhere to great advantage.

Such informal examinations are extremely helpful in the teaching of reading. There follows an illustrative application.

Divide the Pupils of the Fourth-Grade Class (Table 8) into Two Groups of Equal Reading Ability.—Motivation can be increased by dividing the pupils in the illustrative fourth-grade class into two groups of equal ability in reading. This division should be made on the basis of the scores made in the initial test or tests. A division made on the basis of scores on the Thorndike-McCall Reading Scale will prove sufficiently accurate for the purpose.

To make such a division arrange the initial scores on the reading scale in order of size, i. e., largest score first, second largest score second, and so on. Put the ablest pupil into

Group I, the second ablest into Group II, the third ablest into Group II, the fourth ablest into Group I, the fifth ablest into Group I, the sixth ablest into Group II, and so on for the other pupils. This will give two groups of practically identical ability.

The two groups so formed should each be encouraged to select a captain, to decide upon a name for the team, to make or select a team motto, and to do anything else which the ingenuity of the teacher or the initiative of the pupils can originate to increase interest in the competition. The team organization may be used to motivate spelling bees, composition work and the like.

Give an Informal Reading Test at the End of Each Week.—After each week of instruction, the teacher should give a test which she herself has constructed from reading material in the regular class reader. Occasionally it would be well to use instead a selection from the text-book in history, geography or the like.

This test should be in the nature of a contest between the two groups in the class. In order that the pupils may not sacrifice speed of reading to comprehension of what is read and *vice versa,* the weekly test should measure both speed and comprehension. The following discussion describes a method of constructing such a test which is simple and expeditious, and results in a type of test which saves the teacher the labor of scoring papers, and which is interesting and educative to the children.

The first step is to select from the reader two pages which are fairly representative of the other pages of the book, which are, if possible, unbroken by pictures, and which have been previously read by the pupils, though this last is not absolutely essential. Assume that pages 10 and 11 of the Fourth Reader would be a satisfactory selection.

After making this selection the next step is to formulate twenty *true-false* or *yes-no* questions based upon the content of pages 10 and 11. How to construct such a test has just been described.

The next step is to give the following directions to the pupils:

Take out a pencil and a sheet of paper. . . . Write your name near the top of the sheet. . . . Now take out your reader and turn to page 9. . . . I shall read aloud the last paragraph on page 9. Follow me as I read. Just as soon as I read the last word, turn over the page and continue to read silently without help. Read as fast as you can get the thought. When you have finished reading I shall ask you some questions to see how much you can remember of what you have read. Read both pages 10 and 11. Just as soon as you read the last word on page 11 close your book, look at the blackboard, and copy on your paper the number you see there.

Regulate the speed of reading so that the last word on page 9 will be read just as the minute hand of the watch is at some ten-seconds point. At the expiration of the first ten seconds write the number 10 on the blackboard. At the expiration of 20 seconds erase the 10 and write 20. Continue similarly until all the pupils have finished.

When all have finished and closed their books give the pupils the following directions:

Write the numbers 1 to 20 inclusive down the left hand margin of your blank paper. . . . I shall ask you some questions about what you have just read. Each question should be answered with a YES *or a* NO. *If you do not know the answer to any question guess at it. If the answer to the first question is* YES, *write* YES *after number 1. If it is* NO, *write* NO *after number 1. Treat the other questions in the same way.*

If the pupils are unable to write they may be instructed to make a check mark when the answer is *yes* and a cross mark when the answer is *no*. Very young pupils will need a preliminary test the sole purpose of which is to teach them how to take the test.

Read each question aloud, slowly and distinctly, giving each pupil enough time to write his answer.

When the pupils have finished this test from memory have them turn their sheets over and write the numbers 1 to 20 again. Then have them open their books to pages 10 and 11. Read the questions once more to see how many questions the pupils can answer with their books open.

Call the correct answers, once for each test. Have the

pupils score themselves or each other's paper or both and compute their own scores. Have pupils mark answers that are wrong. Omitted questions are counted as wrong. The score on each test for each pupil is found by subtracting two times the number of errors he makes from the total number of questions. Thus if pupil A had six mistakes on the first part and three on the second part of the test, his comprehension scores would be:

Memory comprehension score $= 20 - 2\ (6) = 8$
Visual comprehension score $= 20 - 2\ (3) = 14$

Next, tell the pupils the number of words on the two pages and have each pupil compute the number of words read per minute. This is the pupil's speed-of-reading score.

Have each pupil preserve a record of his own scores from week to week in order that he may see whether he is improving in his speed of reading in particular. This measure of improvement will be necessarily crude because the material of the different tests will vary somewhat in difficulty from time to time.

Compute the mean memory-comprehension score, the mean visual-comprehension score and the mean speed-of-reading score. The teacher should keep a record of these three class scores and observe whether pupils are making progress under her instruction.

The scores made by a small fourth-grade class on such an informal silent reading test are shown in Table 13.

How to Utilize Results.—The scores of Table 13 suggest the following useful conclusions:

1. Pupils A, E, and H are poor readers. Pupil E could do nothing whatever. The three are deficient in every respect due to insufficient training, or improper training, or lack of native intelligence or some other cause. Whether it is due to insufficient training can be determined by watching the progress when further training is applied or by enquiring concerning the child's educational history. Whether it is due to improper training can be determined by measuring progress after giving proper training.

TABLE 13

Scores Made by a Fourth-Grade Class on an Informal Silent Reading Test

Pupil	Memory Comprehension	Visual Comprehension	Speed
A	2	6	125
B	5	8	127
C	7	12	150
D	5	16	121
E	0	0	...
F	4	18	200
G	7	10	150
H	0	2	100
I	13	16	131
J	6	12	146
K	8	10	150
L	18	20	180
M	15	16	172
N	10	12	146
O	9	12	150
P	4	8	104
Q	10	14	140
R	7	12	140
S	12	14	166
T	6	10	127
Mean	7.4	11.4	136.3

Whether it is due to lack of sufficient native intelligence is best determined by means of an intelligence test.

2. Pupil D has an exceptionally poor memory. That he can comprehend what he reads is shown by his high score at visual comprehension. The slow rate at which he reads indicates that his poor showing was not due to hasty reading.

3. Pupil F makes an exceedingly low score at memory comprehension, considering his very high record at visual comprehension. In all probability this is not due to a native deficiency of memory, but to extremely rapid reading. In fact, pupil F reads more rapidly than anyone else in the

class. He should be taught how to control his speed of reading.

4. As indicated by the memory-comprehension and visual-comprehension scores, pupil I is a very superior reader. However, he is not a particularly speedy reader. Above all he needs training in speed.

5. Pupils L and M are superior readers in every respect. They read rapidly. They have a tenacious memory. Their visual comprehension, which is probably the most important of the three, is equally superior.

6. Records not included in Table 13 are also useful. As previously recommended have those pupils who answer the first question correctly on the last part of the test, hold up their hands. Do this for each question. Questions missed by many in the class should be discussed. Finding the questions which have been missed by the class reveals the place where teaching is needed.

Give an Informal Test of the Speed and Quality of Oral Reading.—William S. Gray has done more careful experimentation with methods of measuring the speed and quality of oral reading than anyone else. The teacher cannot do better than follow the procedure which he has evolved. This procedure is modified below to fit an informal test.

He suggests that each pupil be tested individually in some quiet place free from distractions and where other pupils cannot hear what is taking place and thus profit, when their turn comes, by the mistakes of their predecessors.

The teacher should hand to the pupil his regular reader open at a previously selected page where the child is to read. As she does this she should say: *I want you to read page .... aloud for me. Begin with the first word at the top of the page when I say Begin! Read until I tell you to stop. If you find some hard words, read them as best you can without help and continue reading.*

In case a pupil hesitates several seconds on a difficult word, pronounce it for him and mark it as mispronounced. If the first word at the top of the page is not the beginning

of a sentence begin time and error records the instant the pupil completes the partial sentence.

Record the exact second when the pupil begins the first whole sentence and the exact second when the pupil finishes the page, paragraph, or whatever amount of material the teacher has previously elected to have read. The number of words read per minute is the pupil's speed-of-oral-reading score.

While the pupil is reading the teacher should carefully record the errors made. The following, quoted from Gray, illustrates the character of the errors and the method of recording them.

> The sun pierced into many my large windows. It was the opening of October, and the clear sky was of a dāzzling blue. I looked out of my window and down the street. The white houses of the long, straight street were almost painful to the eyes. The clear atmosphere allowed full play to the sun's brightness.

"If a word is wholly mispronounced, underline it as in the case of 'atmosphere.' If a portion of a word is mispronounced, mark appropriately as indicated above: 'pierced' pronounced in two syllables, sounding long *a* in 'dazzling,' omitting the *s* in 'houses' or the *al* from 'almost,' or the *r* in 'straight.' Omitted words are marked as in the case of 'of' and 'and'; substitutions as in the case of 'many' for 'my'; insertions as in the case of 'clear'; and repetitions as in the case of 'to the sun's.' Two or more words should be repeated to count as a repetition.

"It is very difficult to record the exact nature of each error. Do this as nearly as you can. In all cases where you are unable to define clearly the specific character of the error, underline the word or portion of the word mispronounced. Be sure you put down a mark for each error. In case you are not sure that an error was made, give the pupil the benefit of the doubt. If the pupil has a slight foreign accent, distinguish carefully between this difficulty and real errors."

The pupil's quality-of-oral-reading score is found by multiplying the number of errors made by 100 and by dividing this product by the number of words in the passage read. A large score should be interpreted as poor oral reading.

The speed-of-oral-reading score for the class is the mean of the speed-of-oral-reading scores for the individual pupils, and the quality-of-oral-reading score for the class is the mean of the quality-of-oral-reading scores for the individual pupils. Both these class scores and the pupil scores should be preserved in order to measure the amount of growth.

III. The Use of Standardized Scales

8 and 9. ***Determine the Final Reading Age and Mental Age.***—The final reading age for the end of the current year must be estimated and so must the mental age. In addition to its other valuable functions the Intelligence Quotient aids us in estimating what each pupil's reading age or mental age should be at the end of, say, ten months of instruction, for a pupil's I.Q. is a rather accurate index of his capacity for progress in silent reading. Reading Quotient will serve as a reasonably good substitute when I.Q. has not been determined. A pupil whose I.Q. is 90 should be expected to progress only 90 per cent as fast or as far in 10 months as a pupil whose I.Q. is 100. Hence to estimate a pupil's final reading age all that is necessary is to add to his initial reading age 90 per cent of 10 months if the pupil's I.Q. is 90 and the school term is 10 months, or 90 per cent of 8 months if the pupil's I.Q. is 90 and the school term is 8 months. Thus pupil A's final reading age is 121 plus 100 per cent of 10 months, i. e., 131, as shown in line 8 of Table 8. His final mental age, computed similarly, is 121 plus 100 per cent of 10, i. e., 131, as shown in line 9 of Table 8. Pupil B's final reading age is 130 plus 122 per cent of 10, i. e., 142. His final mental age is 150 plus 122 per cent of 10, i. e., 162. These computations assume that the school term is 10 months.

I.Q. will prove a reasonably satisfactory basis for estimating progress in all such complex functions as silent reading. It may not prove satisfactory for such a narrow function as handwriting. Prophecy as to progress in handwriting may have to be based upon a measurement of special capacity rather than general intelligence. Lacking this the teacher may use Handwriting Quotient or may arbitrarily assign objectives.

10. ***Determine the Objective for the End of the Year.***—The objective in reading for each pupil should be either his estimated final reading age or his estimated final mental age. As a rule it will doubtless prove more just and satisfactory both to teacher and pupil to use the estimated final reading age as the minimum objective in preference to the estimated final mental age. The reading age as an objective requires each pupil to make normal progress for his capacity. If he has, in the past, exceeded normal expectation, this objective will preserve all such gain. Mental age on the contrary would not preserve the gain. Furthermore, any pupil whose Accomplishment Quotient was below 100, as for example pupil B, would be required to do the difficult task of making up all this deficiency in a single year. Finally mental age would place a heavy burden upon schools whose term is short. Hence the estimated final reading age for each pupil should define the minimum objective in reading.

But the estimated final reading age should be considered a minimum and not a maximum objective. It makes no provision for recovering past losses on the part of pupils whose Accomplishment Quotients are less than 100. These pupils in particular should look upon their objectives as low minima. Reading age objectives should not be stopping places for any pupil no matter what his Accomplishment Quotient. Such objectives are meant to be passed and, if possible, left far in the rear. Any teacher who faithfully carries out the procedure here outlined should expect to exceed such goals. There should be minimum objectives,

but there should not be any maximum objectives except in the sense that other purposes and abilities of vital importance should not be sacrificed to attain some far-off goal in reading. However, it should not be forgotten that he who increases his ability in reading is contributing to an increase in many other abilities.

The objective in terms of reading age should be translated back into a score on the reading scale. Pupil A's estimated final reading age is 131. By consulting Table 11 it is found that 131 corresponds to a T score of 43. Hence pupil A's minimum objective is 43. Pupil B's objective is 47. The objectives for the other pupils are shown in line 10 of Table 8.

The objective for the class as a whole is the mean of the score objectives for the individual pupils. The class objective is then, as shown, 39.8.

The common practice of setting up no definite visible objective at all could not be expected to produce other than the current indifference toward improvement. We would question the intelligence of any adult who seemed to be in a great hurry if he did not know where he was nor where he was going. Without the initial measurements already recommended, children, as Foote of Louisiana points out, practically do not know where they are and without more definite objectives they do not know in any thrilling way just where they are to go. It may be a tribute to children's intelligence that they are listless and uninterested.

The common practice of setting up definite objectives which are not objectives at all but impossible ideals for the class can only produce discouragement. Either because of delusions of grandeur concerning their own efficiency or because of an irrational confidence in their pupils non-technically-trained teachers and supervisors almost invariably set up impossibly distant objectives. Recently a group of unusually progressive teachers decided to set up objectives in composition. After months of study sample compositions were selected to mark the passing point for each grade.

When these specimens were measured on a standardized composition scale it was found that the specimen selected to indicate the passing point for the fifth grade was of a quality which twenty-five per cent of sophomore college students could not equal.

The common practice of setting up a definite objective which is reasonable for the class as a whole, but which is the same for all the pupils in the class is almost equally bad. It violates a fundamental psychological law that pupils differ and differ greatly in both their initial ability and their capacity to make progress. The following fable from the *Rose Garden* of the Persian poet, Sa'di, the "nightingale of Shiraz," is still true:

> "A king handed over his son to a teacher and said, 'This is my son; educate him as one of thine own sons.' The perceptor spent some years in endeavoring to teach him without success, while his own sons were made perfect in learning and eloquence. The king took the perceptor to task, and said, 'Thou hast acted contrary to thy agreement, and hast not been faithful to thy promise.' He replied, 'O King! education is the same, but capacities differ.'"

Graph the Immediate and Final Objectives.—It is well to have not only annual but also monthly objectives for each pupil and to have these graphed by the pupils themselves. There are enough of the Thorndike-McCall Reading Scales to give one each month, consequently each pupil should divide the total distance he must travel during the year into ten equal portions if the school term is ten months, or five equal portions if the school term is five months. Pupil A's initial score is 40. His final objective is 43. He must gain 3 points in 10 months, i. e., three-tenths of a point each month. A large sheet of coördinate paper may be tacked upon the wall of the school room and each pupil can lay off his monthly objectives on this single large sheet.

Graph the Objectives for All Classes in the School.—Progress in reading can be motivated even better if all the classes in the school are carrying out the reading program outlined. In this case, or even when just the adjoining grades participate, it would be advisable to construct a large

chart showing the immediate and final goals for each grade. This chart should be posted in some conspicuous place in the school building. The classes may then compete to see which one can first attain its minimum objective. The monthly progress made by each class should be indicated on this school graph.

The particular fourth grade shown in Table 8 has an initial score of 36.4 and a final objective of 39.8. Hence for this class the school chart would graphically show the initial point and the subsequent monthly hurdles, thus:

36.4 36.7 37.1 37.4 37.8 38.1 38.4 38.8 39.1 39.5 39.8

Lest the graphic location of pupil objectives on a publicly posted chart humiliate any pupil who saw that his bar on the graph was shorter than other bars, the teacher should make the diagram in such a way that all bars are of equal length. The principal of the school should follow a similar plan in constructing the diagram for the contest between classes. The teacher and principal would know that one inch, say, of a bar meant, perhaps five points for one child or class and, possibly, eight points for some other child or class.

Visible Goals for Education.—During the World War an indispensible device for increasing military efficiency was a battle map showing the present positions of the contending armies and clearly defining the objectives of attack. Similarly all great industrial corporations find it necessary to keep a production chart which shows the extent to which actual production has kept pace with the desired production. Education is the world's greatest manufacturing industry. Completely surrounded by its foes it is attacking in all directions in the world's greatest battle. And yet education has neither battle map nor production chart. Its goals can only be made visible as objective measurement develops.

To carry the battle analogy further, the attack of the school should be like the attack of the various divisions of a large army. The General Staff marks out a series of goals for each division. Certain divisions move forward and capture their objectives before certain other divisions even leave

their trenches. When a division has attained its first objective it is instructed to "dig in" until other divisions have reached their assigned goals. To move divisions forward without regard to the position of coöperating divisions would mean speedy disaster. But pupils are tough. They are frequently pushed limpingly toward one goal before they have reached a prerequisite or more valuable goal in another goal series. Such neglect of proper emphasis will continue until objective measurement has made visible both the curriculum of purposes and the curriculum of abilities as well as the position of the pupil with reference to the curriculum.

This suggestion that tests be used as the regulator of educational emphasis will be opposed by a large group of well-meaning educators. The humanity of Pestalozzi and the sympathy for childhood of the good people who have followed in his train could not abide the dry-as-dust drill to which children were subjected. The reaction away from the drill subjects by certain educators is more than an emotional one. It is in part due to a real change in their conceptions of what is most worth while in education. They desire, and rightly so, a greater emphasis upon those virtues which have to do with civic responsibilities and other relationships. Since most existing tests measure drill subjects there is a grave fear that the widespread use of tests will merely increase the emphasis upon what they conceive to be the relatively less valuable abilities.

The attitude of these educators is wholly honest but substantially unwise. There is grave danger that they will use their ingenuity not to devise ways of tying the skills to children's purposes in such a way that drill will be interesting, but to undermine our conception of the tremendous value of these skills. Tom dreamt that he and many other chimney sweeps were locked in black coffins and that there came an angel with a golden key and unlocked the coffins and set them all free. The golden keys with which teachers unlock the minds of children are the basic skills. They are more valuable even than Virgil's golden bough for they open

the very gates of life. The skills are valueless in themselves, and at the same time they are the indispensable prerequisites of all that is valuable in education. Like the centaur's tunic they cannot be torn off without carrying away the flesh and blood of the wearer.

Take reading, for example. Carlyle was not far wrong when he said that all any school can do is to teach us how to read. Carlyle tells how Odin was credited with the greatest invention man has ever made, namely, the invention of letters whereby man may mark down the unseen thoughts that are in him. He tells of the astonishment of Atahualpa the Peruvian king; how he made the Spanish soldier who was guarding him scratch *dios* on his thumb-nail, that he might try the next soldier with it and thus ascertain whether such a miracle were possible. Odin deserved his deification.

A curriculum contains but two absolute indispensables. These are purposes and those particular types of abilities called skills. Purposes are unquestionably primary; skills are next; information and similar types of abilities are least important. The importance of purposes has already been emphasized. Those who are asking for a different emphasis in education are really asking for purposes. Most of the abilities listed as being the higher values of education, really are purposes. Most of the problem of what is called socializing pupils reduces to a problem of inculcating purposes. Most individuals possess sufficient ability to be honest, courteous, sympathetic, coöperative, generous, unselfish, and all the rest of the Christian virtues. Their lack is not ability but purpose.

Skills are a close second to purposes in worth. Skills are methods of work. How to manipulate numbers, how to write, how to compose, how to read, how to use books so as to find a bit of desired information, how to evaluate material, how to think, etc., all these are skills which match purposes in worth. The pupil who has learned how to work and

learn, and who is provided with a rich set of purposes may be turned loose to educate himself.

Information is least in importance. Most of what all of us once learned, we have forgotten without regret. But not for anything would we give up our purposes and our skills which enable us to learn new information or relearn the old when needed. Many skills can only be developed by exercising them upon knowledge, facts, or information. The materials of information and the like selected for whetting the mental skills should be those which will be most serviceable for the realization of purposes. But all the time knowledge should, from the adult's point of view, be considered merely the by-product of the process of developing purposes and skills.

Instead of objective tests causing an over-emphasis upon the skills to the detriment of purposes, their use will insure to skills just that emphasis provided by the curriculum and will prove to be the salvation of the higher values. When intelligently used tests are merely instruments for realizing the curriculum. Like poison, steam engines, fire or any other potent force they require intelligent control. We do not trust fire to infants, and if there exists anywhere educators who do not subordinate tests to their curriculum, they are still in their professional infancy and should not be trusted to use tests.

Tests will be the salvation of the higher values because of a natural human tendency to stress the tangible and visible. Just as a child will not put forth intense effort when he can see no results, so a teacher is not likely to spend much effort trying to develop a trait improvement in which neither she nor the child's parents can see. When a month's improvement in handwriting or composition is invisible, a year's improvement in unselfishness, even though very important will scarcely tip the scales of consciousness. It is human nature to fix our faith to form, hence so long as the average of human nature remains what it is, we must not expect it

to expend effort in producing invisible, unrewardable improvements so long as it is permitted to produce visible rewardable changes. The moon pulls on the earth as well as on the sea but the earth tides interest few. Visibility and rewardability control the amount and direction of effort. The skills have been over-emphasized in the past, and always will be until we have either the thus-far-and-no-farther of tests or an educational magnifying glass which will make visible what has before been invisible. Even though "We are such stuff as dreams are made of" we simply refuse to "Pipe to the spirit ditties of no tone."

Give the Standard Reading Scale at the End of Each Month.—The teacher should give the Thorndike-McCall Reading Scale monthly. Pupils should be encouraged to look upon these monthly tests as important events. Each pupil and class should strive to exceed the next objective. After the tests have been scored, the results should be officially graphed upon the pupil and class charts.

When a pupil or a class fails to show progress both teacher and pupils should attempt to locate the cause. In a school where this program is now in operation the third grade, at the end of the first month, exceeded its goal for the end of the year. The fourth grade progressed three-fourths of the distance toward its goal for the end of the year. The fifth grade made no progress whatever. The sixth and seventh grades made excellent progress. The eighth grade, like the fifth, made no progress whatever. Such startling differences as these call for careful investigation. Due to crude scoring units or to the unreliability of the test individual pupils will appear frequently to have made no progress, but this does not explain the lack of progress of a whole class.

CHAPTER V

MEASUREMENT IN EVALUATING THE EFFICIENCY OF INSTRUCTION

11, 12 and 13. ***Determine the Final Accomplishment Quotient.***[1]—The score on the last standard test of each pupil should be converted into a reading age, and this actual final reading age should be divided by the estimated final mental age. The quotient is the final Accomplishment Quotient which, taken in conjunction with the initial Accomplishment Quotient, is the final evaluation of the efficiency of the year's work.

Line 11 of Table 8 shows that pupil A made a final score of 43, which, as line 12 indicates, is equivalent to a final reading age of 130. Line 9 shows him to have an estimated final mental age of 131 months. Dividing his final reading age by this mental age gives him an Accomplishment Quotient of 99. Thus pupil A has substantially attained his objective. Pupil B's Accomplishment Quotient markedly improved during the year, thereby showing that he greatly exceeded his minimum objective. Nevertheless his achievement is still considerably below what it should be. Pupil C not only maintained his high initial Accomplishment Quotient of 112, but pushed it 6 points higher.

The mean initial reading score for the entire class was 36.4. The reading score objective was 39.8. The actual final reading score made was 41.2. Thus the class considerably exceeded its minimum objective. The mean initial Accomplishment Quotient was 100. The final Accomplishment Quotient was 103. Irrespective of what the contest

[1] Those portions of Chapters III, IV, and V which have particular reference to reading are brought together in somewhat modified form to make a chapter in the Teachers' Manual which accompanies *The Child's World Readers*, published by Johnson Publishing Company, Richmond, Va.

between classes revealed (results not presented here) both the teacher and pupils of our illustrative fourth-grade class know that they have a right to be proud of the record made.

Rate the Efficiency of Teaching.—Do the results from standard tests given to a class reveal the efficiency of the teacher of that class? They do and they do not. They do provided certain conditions obtain. These conditions are roughly obtainable by an experimental control of the situation. They do not, because the conditions necessary for a just evaluation of a teacher's efficiency rarely obtain in the ordinary uncontrolled testing situation.

All told more harm may easily be done than good. The use of tests for the guidance and diagnosis of pupils is so much more vital than their use to evaluate teachers that the former value should not be lost through antagonizing teachers in order to obtain the latter. For some time to come, at least, tests had better be used to measure pupils and not teachers, except in so far as teachers measure their own efficiency or coöperate in its measurement. When tests have reached a state of development where their use will lead to a just evaluation, the really efficient teachers will themselves demand to be rated by means of tests in order to escape another method whose accuracy is such that educators tolerate it only because nothing else has been available.

Since, however, teachers and supervisors are both likely to demand in the near future that their work be evaluated in a more scientific and hence more impersonal manner, there is summarized below what I conceive to be the fundamental assumptions underlying a scientific procedure for rating and promoting teachers and supervisors as well as the steps in the process of making such ratings.

1. The pupil is the center of gravity or sun of the educational system. Teachers are satellites of this sun and supervisors are moons of the satellites.

2. All the paraphernalia of education exist for just one purpose, to make desirable changes in pupils.

3. The worth of these paraphernalia can be measured in just one way, by determining how many desirable changes they make in pupils.

4. Hence the only just basis for selecting and promoting teachers is the changes made in pupils.

5. Teachers are at present selected and promoted primarily on the basis of their attributes, such as intelligence, personality, physical appearance, voice, ability in penmanship and the like.

6. No one has demonstrated just what causal relationship, if any, exists between possession of these various attributes and desirable changes in pupils. The relation between possession of certain attributes and the degree of favor of a teacher in the inspector's eyes is more evident. Dr. Chassell in her Ph.D. thesis determined the correlation between certain features of Ph.D. students of education and later success. She found that the score made in Ph.D. matriculation examinations at Teachers College correlated with success about .50. The quality of their Ph.D. dissertations correlated about .50. The letters of recommendation written about these Ph.D.'s correlated about .30. Their handwriting correlated .20, and their photograph .10. The following showed substantially zero correlation with later success: physical defects, type of locality of birthplace, age of reaching a given academic status, study abroad, size of family, church relationship, reading knowledge of languages, and travel abroad. This study is more valuable in the present connection for the technique it exemplifies than for its conclusions. The subjects were not typical teachers but Ph.D.'s. The criterion of success was not demonstrated changes in pupils but the opinion of judges.

7. Scientific measurement itself is fair only when we measure the amount of desirable *change* produced in pupils by a given teacher. The measurement of *change* requires both initial and final tests. The plan outlined below provides for these.

8. Scientific measurement is fair only when we measure amount of change produced in a *standard time*. This requirement can be satisfied.

9. Scientific measurement is fair only when we measure the amount of change in *standard pupils*. The Accomplishment Quotient is included in the plan below because this is a device for converting pupils, no matter what their intelligence, into standard pupils.

10. Scientific measurement is fair only when the measurement is complete. Absolute completeness would require a measurement of the amount of changes made in children's purposes as well as their abilities. Absolute completeness is of course impossible and is in fact not necessary; partly because a chance sampling of the changes made will be thorough enough, and partly because teachers' skill in making desirable changes in, say, reading is probably positively correlated with their skill in making desirable changes in, say, arithmetic.

The technique for satisfying the foregoing requirements and evaluating a teacher's efficiency in, say, reading follows:

1. Determine the initial reading score of the pupils in the teacher's class (line 2 Table 8).
2. Convert initial reading score into reading age (line 3 Table 8).
3. Determine each pupil's mental age at the same time the initial reading test is given (line 5 Table 8).
4. Divide mental age by chronological age to get I.Q. (line 6 Table 8).
5. Divide reading age by mental age to get A.Q. in reading (line 7 Table 8).
6. Estimate final mental age for the end of the teaching period (line 9 Table 8).
7. Determine final reading score at the end of the teaching period (line 11 Table 8).
8. Convert final reading score into final reading age (line 12 Table 8).

9. Divide final reading age by estimated final mental age to get final A.Q. (line 13 Table 8).

10. Subtract mean initial A.Q. in reading from mean final A.Q. in reading (line 14 Table 8).

If the mean difference is zero the teacher has been typically efficient. If the difference is below zero the teacher is below average in efficiency. To the extent that the mean difference is above zero, just to that extent the teacher has shown superior efficiency in teaching reading.

11. Repeat steps 1, 2, 5, 7, 8, 9, and 10 for arithmetic and the mean A.Q. difference will be an index of the teacher's efficiency with arithmetic. In similar fashion, her efficiency at teaching other measurable abilities could be determined.

12. Compute the mean of these various mean A.Q. differences to get a final determination of the teacher's efficiency. Franzen has defined a teacher's efficiency in terms, then, of the following formula where N is the number of subjects tested or tests given.

$$\text{Teacher Efficiency} = \frac{(\text{Read. A.Q. Diff.}) + (\text{Arith. A.Q. Diff.}) \text{ etc.}}{\text{N}}$$

In the case of our illustrative fourth-grade class the teacher's efficiency was as follows:

$$\text{Teacher Efficiency} = \frac{103 - 100}{1} = +3$$

This formula may be used not only for the rating of teachers but also for selecting new teachers who are given a half-year's or year's trial.

Crude as the proposed method of selection is, it is fairer than present methods. The superintendent doubtless soliloquizes like Caliban upon Setebos as the applicants march before him with a sample of painstaking penmanship in one hand and an antique photograph in the other.

> "Am strong myself compared to yonder crabs
> That march now from the mountain to the sea;
> Let twenty pass and stone the twenty-first,

Loving not, hating not, just choosing so.
Say, the first straggler that boasts purple spots
Shall join the file, one pincer twisted off;
Say, this bruised fellow shall receive a worm,
And two worms he whose nippers end in red;
As it likes me each time, I do: so He."

Rate the Efficiency of Study.—Henry is a relatively stupid boy but his father doesn't know it. The teacher doesn't know it. The teacher considers him lazy. Two years ago Henry was in a class with pupils of his own age. Owing to his low intelligence he was hopelessly outclassed and as a consequence was failed by his teacher. When the father received the report, he and Henry had a dramatic session in the woodshed.

Henry repeated the work of the grade with another class which happened to be younger and stupider than usual. As a result of this fortunate combination in his competitors, Henry's father received at the end of the year a good report of Henry's work. Henry's teacher is happy because she thinks she succeeded so much better with him than did his former teacher. The former teacher is happy because she thinks it was her courage in failing him that paved the way for a moral reformation. Henry's father is happy because he considers he knew just exactly the right stimulus to use to motivate Henry's study. Henry is happy because he is not as unhappy as he was a year before.

This year Henry is fighting a losing battle in competition with those who are intellectually superior. He already sees that he is headed straight for another failure which does not worry him, and another thrashing which concerns him greatly. Thus every other year Henry will receive his inevitable thrashing until he is strong enough to physically rebel.

Henry is not one child but a million children in this land of justice. These million are yearly subjected to such injustice because reports sent to parents are misleading.

The intellectually superior children suffer as much as or more than stupid children, but their suffering is of a different type. Most gifted children are working far below their

optimum level of efficiency simply because no one suspects their real possibility. Since they lead their classes without difficulty and hence secure all the rewards which additional effort would bring there is no motive to exceed their present rate of progress.

How may a just report be made? The Accomplishment Quotient not only measures the efficiency of a whole school but it also yields the fairest measure of the extent to which a pupil has progressed in proportion to how much he was capable of progressing. Hence the Accomplishment Quotient is at least one of the measures which should be sent to parents.

A measurement of the efficiency of pupils is likewise useful in conferences with parents. Consider for a moment how much more useful a principal could make himself if he possessed for every pupil in his school the information shown in Table 2. Presented with an array of such impartial facts, the parent who came to scoff would remain to pray. Parents who came earnestly seeking means to coöperate would not go away empty of fruitful suggestions.

Fortified with such information the principal would be equally useful in conferences with teachers and pupils. Given such a detailed knowledge of the conditions in the school and of the problems with which each teacher is contending, the principal could add to his teacher's respect for his superior power, a respect for his superior knowledge. As the situation now stands the average principal must honestly confess that the rank and file teachers know far more than he of the real condition of the school. Finally there would be innumerable instances where such information would enable him more intelligently to confer with pupils, to deal with discipline cases, and to supervise the instruction of individual children.

Experimental Selection of Methods and Materials of Study, Instruction, and Supervision.—Here stands a pupil—the Alpha and Omega of all educational effort, the center of gravity of the educational universe. Everything

that Midas touched turned to gold. Everything that touches a pupil shows whether it is gold. Teacher, supervisor, principal, superintendent, United States Commissioner of Education, materials, methods, normal schools, this book, educational tests, the educational philosopher who confines himself solely to a contemplation of the ultimate, all these show whether they are gold or dross by the efficiency they show in altering the synaptic connections of this pupil's neurones. If no one of the above produces any desirable change in the pupil they are educationally without worth. Educational measurement is distinctive in that it must show the educational efficiency of all things and then in the last great experiment show whether it too has or has not value. Thus measurement alone possesses the power of self-destruction. And its worth like the worth of all else depends upon the amount and value of the changes it can produce in this pupil.

If probability may be so crowded as to assume that all of the above have an educational value above zero, the next questions become: Which of two methods is more efficient? Which of two teachers or text-books is more efficient? Which of two practice tests? Which of two forms of organization? As has previously been suggested, an absolutely just answer to each of these questions requires a carefully controlled scientific experiment.

Space will not permit more than a listing of the ideal conditions aimed at by one of the simplest types of educational experiments, namely, the equivalent groups experiment.

1. Two groups of pupils which are absolutely equivalent as shown by absolutely accurate and adequate initial measurements. The two groups must be more than equivalent as to averages. Every pupil in one group must be paired by an equivalent pupil in the other group. Increasing the size of the group will usually aid in securing equivalence. While in theory the initial measurements should be absolutely thorough, in practice they are seldom more than an accurate and adequate measurement of those abilities which

the methods or materials or whatever is being contrasted are expected to alter.

2. Except for the experimental factors the maintenance of absolutely identical conditions for the two groups for the entire period of the experiment. If the purpose of the experiment is to decide which of two methods of teaching silent reading is the more effective, the conditions surrounding each group are kept identical except that Group A is taught silent reading by Method A and Group B is taught by Method B.

3. The maintenance of each experimental factor at exactly the desired intensity.

4. An absolutely accurate and adequate measurement of the final ability of each group of pupils.

5. A thoroughly just evaluation of the total worth of all changes occurring in Group A as compared with Group B.

6. A final conclusion which is formulated in the light of the type of pupils used as subjects, and the intensity of each experimental factor.

7. A statement of the statistical reliability of the conclusion, if there is the least suspicion that the ideal requirements have not perfectly obtained. In actual practice this, of course, means that the statistical reliability of the conclusions should always be stated.

Efficiency Measurement of Schools and School Systems.—In extensive surveys of schools and school systems the careful methods of evaluating efficiency described in the preceding pages must usually give way to a cruder and more rapid method. Such a rapid method is illustrated in Table 14.

Table 14 is read thus: On the Thorndike Reading Scale Alpha 2, the 1919 median for Grade III is 5.5. The norm is 8.0. The difference between the 1919 median and norm is a negative 2.5. The grade unit is negative 0.6. Stated differently, Grade III is below norm 2.5 which is equivalent to being below norm 0.6 of a grade. Grade IV is below norm one entire grade. Grades V, VI, VII, and VIII are

TABLE 14

Comparison of the Efficiency of School X for Each Test for Each Grade with Normal Efficiency. (Data from Table 2.)

	III	IV	V	VI	VII	VIII	All Grade Mean	VI-VII-VIII Mean
Thorndike Reading Scale Alpha 2								
1919 Median	5.5	10.8	18.0	23.0	24.6	25.9		
Norm	8.0	15.0	20.0	24.0	28.0	30.0		
Difference	— 2.5	— 4.2	— 2.0	— 1.0	— 3.4	— 4.1		
Grade Unit	— 0.6	— 1.0	— 0.5	— 0.2	— 0.8	— 0.9	— 0.7	— 0.6
Trabue-Kelley Completion								
1919 Median	16.5	29.0	38.5	42.5	48.3	55.5		
Norm	18.5	25.0	30.5	35.0	40.0	43.5		
Difference	— 2.0	4.0	8.0	7.5	8.3	12.0		
Grade Unit	— 0.4	0.8	1.6	1.5	1.7	2.4	1.3	1.9
Thorndike Vocabulary Scale A_2x								
1919 Median	21.5	51.5	81.5	106.5	108.5	109.5		
Norm	30.0	65.0	83.0	95.0	108.0	117.0		
Difference	— 8.5	— 13.5	— 1.5	11.5	0.5	— 7.5		
Grade Unit	— 0.5	— 0.8	— 0.1	0.7	0.0	— 0.4	— 0.2	0.1
Ayres Spelling Test								
1919 Median	13.5	21.5	30.5	47.5	48.3	55.5		
Norm	19.6	30.4	37.8	47.7	50.3	54.4		
Difference	— 6.1	— 8.9	— 7.3	— 0.2	— 2.0	1.1		
Grade Unit	— 0.9	— 1.3	— 1.0	0.0	— 0.3	0.2	— 0.6	0.0
Woody Addition Scale, Series B								
1919 Median	7.1	12.6	12.0	14.0	15.5	15.8		
Norm	9.0	11.0	14.0	16.0	18.0	18.5		
Difference	— 1.9	1.6	— 2.0	— 2.0	— 2.5	— 2.7		
Grade Unit	— 1.0	0.8	— 1.1	— 1.1	— 1.3	— 1.4	— 0.9	— 1.3
Woody Subtraction Scale, Series B								
1919 Median	0.7	7.3	8.8	10.0	12.2	12.8		
Norm	6.0	8.0	10.0	12.0	13.0	14.5		
Difference	— 5.3	— 0.7	— 1.2	— 2.0	— 0.8	— 1.7		
Grade Unit	— 3.1	— 0.4	— 0.7	— 1.2	— 0.5	— 1.0	— 1.2	— 0.9
Woody Multiplication Scale, Series B								
1919 Median	3.9	7.5	8.7	10.7	14.1	12.5		
Norm	3.5	7.0	11.0	15.0	17.0	18.0		
Difference	0.4	0.5	— 2.3	— 4.3	— 2.9	— 5.5		
Grade Unit	0.1	0.2	— 0.8	— 1.5	— 1.0	— 1.9	— 0.8	— 1.5
Woody Division Scale, Series B								
1919 Median	3.4	5.4	4.9	8.0	10.1	9.8		
Norm	3.0	5.0	7.0	10.0	13.0	14.0		
Difference	0.4	0.4	— 2.1	— 2.0	— 2.9	— 4.2		
Grade Unit	0.2	0.2	— 1.0	— 0.9	— 1.3	— 1.9	— 0.8	— 1.4
Nassau County Composition Scale								
1919 Median	2.5	3.0	4.1	4.7	5.0	6.2		
Norm	2.1	2.6	3.0	3.6	4.1	4.8		
Difference	0.4	0.4	1.1	1.1	0.9	1.4		
Grade Unit	0.7	0.7	2.0	2.0	1.7	2.6	1.6	2.1
Mean Grade Unit.	— 0.6	— 0.1	— 0.2	— 0.1	— 0.2	— 0.3	— 0.2	— 0.2

below norm respectively, 0.5, 0.2, 0.8, 0.9 of a grade. All grades average 0.7 of a grade below norm. The last three grades average 0.6 of a grade below norm. The other tests are read in the same way. The grand total mean for each of the last two columns for all tests is in each case 0.2 of a grade below norm.

The only new procedure in Table 14 is the computation of grade units. It is clear that differences between 1919 medians and norms are not comparable from test to test. The difference of 1.4 for Grade VIII on the composition scale is, as shown, equivalent to 2.6 grades while a difference of 1.5 for Grade V on the vocabulary scale is equivalent to only 0.1 of a grade. A difference of 1.4 on the composition scale means much because the total progress of the norm from Grade III to Grade VIII is only (4.8–2.1) or 2.7. A difference of 1.5 on the vocabulary scale means little because the total progress of the norm from Grade III to Grade VIII is (117–30) or 87. Before the differences on the various tests could be combined they had to be made comparable. Comparability was secured by reducing all difference to grade-unit differences.

The computation of the above grade units was as follows: Each difference was divided by the mean amount of norm progress for each grade for the test in question. On the Thorndike Reading Scale Alpha 2 the norm for Grade III is 8.0, for Grade VIII is 30.0. The progress is (30.0–8.0) or 22. The mean progress per grade is (22 ÷ 5) or 4.4. On an average, then, the typical pupil progresses 4.4 points for each grade. This means that if any grade of School 4 is 4.4 below the norm, it is really (4.4 ÷ 4.4) or 1.0 grade or 1.0 grade unit below norm. Thus in Grade III, the difference is negative 2.5, which, divided by 4.4, gives negative 0.6 of a grade unit. The difference for Grade IV is negative 4.2 which, divided by 4.4, gives approximately negative 1.0 grade unit, and so on. The mean norm grade progress for the completion test is 5.0, hence all differences for this test are divided by 5.

The grade units might have been computed in many other ways. The mean grade progress might have been the grade progress of School X instead of the norm. An objection to this is that the meaning of the grade unit would vary with the school being studied. By using norm progress, one school which is one grade unit below norm is equivalent to another school which is one grade unit below norm.

Again, instead of dividing each difference by the same mean, each difference might have been divided by the interval between the two adjoining grades which are nearest to the difference, i. e., the third-grade difference might have been divided by the norm progress of Grade IV over Grade III. But occasionally the norm score for Grade IV is less than or exactly equal to that for Grade III. When the norm scores are equal, the progress is zero. Any difference divided by zero gives an infinite quotient.

A still better method is to divide each difference by the mean of the two or three nearest grade intervals. This would partially avoid a difficulty in the method used by me. My method somewhat exaggerates the differences for the lowest and highest grades, due to the fact that progress is usually most rapid in the lowest grades and least rapid in the highest grades. The procedure of dividing by the mean of the few adjoining intervals was not adopted partly because this makes the computation rather laborious, and partly for other reasons.

The educational surveyor does not always find his data in such form that he can proceed forthwith to compute grade units. I met just such a situation in measuring the reading ability of the white and colored elementary schools of Baltimore. The problem and its solution is presented in Table 15. The Baltimore schools were measured at the close of November, whereas the norm and the achievement of particular school systems were known for the last of June.

To make all scores comparable Baltimore's scores for both white and colored schools were computed forward from November to June, i. e., about seven-tenths of the school year.

later. The process for the white schools was as follows: $(44.9-39.6)\times .7=3.7$. $39.6+3.7=43.3$. $(49.0-44.9)\times .7=2.9$. $44.9+2.9=47.8$ and so on for grades VI and VIII. What the eighth grade will be in June presents a special problem which can be solved by making use of the norm or the achievement of the average school system, thus: $(60.9-58.3)\div(58.3-48.0)=.25$. $(58.1-47.8)\times .25=2.6$. $59.5+(2.6\times .7)=61.2$. By locating the scores used and the results secured in Table 15 the reader will have little difficulty in following the computation and the reason for each step.

TABLE 15

Comparison of the Grade Scores on the Thorndike-McCall Reading Scale for the White and Colored Elementary Schools of Baltimore with Grade Scores for the Average School System and the Mean Achievements of Particular School Systems

Grade		IV	V	VI	VII	VIII
Baltimore white	(Nov.)	39.6	44.9	49.0	54.9	59.5
" "	(June)	43.3	47.8	53.1	58.1	61.2
" colored	(Nov.)	35.4	39.7	42.5	46.9	47.0
" "	(June)	38.4	41.7	45.6	47.0	47.9
Average School System	(June)	*41.8*	*48.0*	*53.7*	*58.3*	*60.9*
Paterson, N. J.	(June)	35.5	40.9	49.0	51.7	53.5
33 Wisconsin Cities	(June)	40.9	47.2	52.6	55.3	58.0
Louisville, Ky.	(June)	39.1	43.6	51.7	59.8	60.7
St. Paul, Minn.	(June)	41.8	46.3	53.5	58.0	62.5
18 Indiana Cities	(June)	49.9	58.9	67.0	68.8	71.5

Table 16 illustrates two ways, grade unit and relative position, which were used for reporting the efficiency of Baltimore's schools. In making out Table 16 only comparable June scores were used.

Interpretations and Functions of Efficiency Measurements.—What may we infer about the efficiency of School X from the data of Table 14? We are in a position to evaluate the achievement of (a) each grade on each test, (b) each grade on all tests, (c) all grades on any test, and (d) all grades on all tests.

TABLE 16

A Table to Interpret the Data in Table 15 When Proper Allowance is Made for the Time of Testing

White Schools						
Grade	IV	V	VI	VII	VIII	Mean
Grades Baltimore is ahead of Average School System	0.3	0.0	—0.1	0.0	0.1	0.1
Rank of Baltimore among six School Systems	2	2	3	3	3	2.6
Colored Schools						
Grades ahead of Average School System	—0.7	—1.3	—1.7	—2.4	—2.7	—1.76
Rank among Six School Systems ...	5	5	6	6	6	5.6

It would be too tedious to list the achievement of each grade on each test. The achievement of each grade on all tests is shown at the bottom of the table. Every grade, without exception, is from 0.1 of a grade to 0.6 of a grade below norm. Grade III and Grade VIII are farthest below norm.

Either the third grade teacher is less efficient than the other members of the teaching staff or else there is some special explanation. Perhaps the third-grade had not begun the study of certain abilities measured by the tests. It is almost a shock to discover that the children know almost nothing about subtraction and yet they are ahead of norm in division. For some reason every fundamental except subtraction has been taught. The teacher was sadly in need of a simple standard test to regulate her educational emphasis. Low mentality of pupils may explain the backwardness of the class, but it does not explain the misplacement of emphasis. To properly interpret the results for Grade III requires a more intimate knowledge of conditions than I

possess. The third-grade teacher may be the most efficient teacher in the school.

The achievement of all grades on each test, shown in the last two vertical columns of the table, brings to light some interesting facts. The school is from one to two grades ahead of standard on the completion scale. This might be attributed to unduly low norms were the results not in harmony with the composition test, which, it is claimed, measures approximately the same ability. On the reading scale, vocabulary scale, and spelling scale, the school is either up to norm or slightly below. In every arithmetic fundamental, without exception, School X is below norm about one grade. If the norm is accepted as a legitimate goal, the school needs to give special attention during the coming year to reading and to the fundamentals of arithmetic.

The achievement of all grades on all tests, the final measure of the school's efficiency, is shown as — 0.2 and — 0.2 at the bottom of the table. The data of the entire table, when thus summarized, shows that the school is two-tenths of a grade below the typical school. This conclusion probably holds not only for the tests used but for all the work of the school, if it were possible to measure and average it.

Education must justify itself in the eyes of the public which it serves. The hour has struck for a testing of every institution in existence. Not even as sacred an institution as the public school can escape the searching investigation of its critical patrons. Those whose business it is to control the distribution of the community's funds are asking more and more that each agency which serves the community *show cause* why it deserves the appropriation requested.

Proof of the school's worth should be as scientific as possible. The time has passed when a gullible public will accept as proof of the school's social value the fact that graduates of the elementary school are socially more valuable than illiterates and that graduates of the high school are worth more to the community than those who left school at

the end of the eighth grade. The only proof that will finally be accepted is verifiable changes in pupils of demonstrable social worth.

The best method yet discovered for measuring changes in pupils in such a way that conclusions may be verified is to employ standard educational tests. They are the friends of both teachers and public. They are a leaven working for an unpolitical scientific evaluation of those who operate the schools. They protect the efficient and expose the inefficient.

Educational tests cannot be safely used to evaluate a school's efficiency without the exercise of ordinary discretion. A school can be compared with its own aims, provided these goals have been wisely located, but not with other schools without conditioning all conclusions upon certain possible factors. A school should never be judged except in the light of these conditioning factors.

One conditioning factor is the permanence of the school population. A few years ago there were in a certain large city many more apartments for rent than there were families to rent them. As a result agents would pay moving expenses and frequently give a bonus of a month's rent to secure tenants. As a consequence of this and other causes, seventy-five per cent or more of the population of certain schools changed every year. Few pupils in the upper grades of many schools have been in any one school during their entire school career or even a large portion of it. In testing the pupils in such a school one is evaluating the efficiency of neighboring schools almost as much as the school being tested. The effect of such impermanence may be to make the efficiency of the school appear too low, too high or just right. School X has a relatively permanent school population.

A second conditioning factor is the intellectual calibre of the pupils. To a certain extent their intellectual calibre may be judged by the general social status of the community. On the whole, the unprosperous, undesirable portions of a

city produce children with an average mentality below that of the children coming from the wealthier sections. Fifth Avenue, New York City, may be more wicked than the lower East Side but it has more intelligence. Determination of intellectual calibre by means of intelligence tests is, however, more accurate than estimates from social status. Pupils in School X, according to estimates from social status, will average below normal intelligence. Whether this estimate is correct is to be checked later by tests. The formula for A.Q. provides a procedure for discounting this important conditioning factor.

Bound up with this is a third conditioning factor, namely, the educative value of the home environment. In some communities the parents probably teach as much as the school, while in the others the home actually discourages study.

A fourth conditioning factor is the amount of chronological retardation or acceleration. How to make allowance for this factor was described in the Directions for Using Thorndike-McCall Reading Scale.

A fifth conditioning factor is the increase in ability which comes from mere maturing. This factor has little significance if all that is desired is a measure of comparative efficiency. If it is desired to determine the absolute improvement produced by the school, any improvement due to mere maturing must be subtracted from the total improvement. The actual separation of the contributions of maturity and those of training has never been done. An elaborate experiment would be required.

Courtis [2] attempted this separation in the survey of the Gary schools. He compared the curve of progress with age in certain mental functions which it is reasonable to assume are increased mainly by mere maturing with the growth curve for mental functions increases in which are the conscious objectives of instruction. When the form of the two curves was similar he inferred that maturity and not specific

[2] S. A. Courtis, *The Gary Public Schools: Measurement of Classroom Products;* General Education Board, New York, 1919.

training should have the credit for the increase in the educational trait. It might be argued with equal force that the mental function which the school was attempting to improve would not have increased at all without specific training, that Nature makes provision for increases in finger length but not for increases in spelling ability, and, finally, that the school whose pupils show growth curves in school traits shaped like maturity curves are schools which have continually and perfectly adjusted their curriculum to increases in the capacity of pupils, if not to the social needs of the community.

A sixth conditioning factor, especially where the efficiency of a teacher of only one or two subjects is being determined, is the transfer of the general training of the school or the specific training of other teachers to the abilities under consideration. Many teachers of special subjects would be chagrined to discover how much of the annual progress of their pupils is due to the general or specific training of their colleagues.

A seventh conditioning factor is the distribution of emphasis. A school whose pupils have a different social and vocational distribution may legitimately elect to emphasize certain subjects more and other subjects less than the typical school. Before selecting the tests for a school to be evaluated it is well to study the emphasis of the instruction as shown by the time schedule, and for the principal and teacher to state the subjects in which they expect the school to make the best showing, the next best and so on to the poorest. Tests can then be selected which will not be especially favorable or unfavorable to the school.

When several tests are used and when they are fairly representative the likelihood of being led into an erroneous conclusion is greatly reduced. Had nothing but the four fundamentals of arithmetic tests been given to School X, for example, the judgment of its efficiency would have been a little too severe, for apparently the school has emphasized composition and related subjects to the detriment of the

fundamentals of arithmetic and reading. The discovery that the school partially compensates an inferiority in certain abilities with a superiority in others, makes us consider whether additional tests might not reveal that the school is up to norm. The fact that reading is below norm does not, however, hold out much hope, for reading is probably the key to more abilities than composition. This is enough to show the importance of remembering to interpret results in the light of the school's emphasis.

There are many other conditioning factors which should not be overlooked. One is the length of the school term. It is not reasonable to expect certain southern rural schools to accomplish in a few months as much as city schools in ten months. The process of allowing for this is too simple to require description. Other factors which must be considered are the training of the teachers, the equipment of the school, and the like.

Dewey has suggested the desirability of a sort of exchange bureau for exchanging educational ideas. Standard tests are a particularly valuable means for communicating ideas and practices. It is one thing for a school to tell what it is doing, and another thing for it to show what it has achieved. Comparisons of the achievement of the schools in one locality with that of schools in other localities will force a discussion of what ability and how much of it is of most worth in education, and thus speed up scientific location of goals. The manifold differences in the achievements, emphasis, and objectives of various schools are not so much indices of a healthy condition as of a fundamental ignorance of the worth of varying amounts of each ability.

Again efficiency comparisons from school to school will hasten the development of a national conception of education. The formulation of a national program of education should be made in the light of what the backward schools of the country are now doing. Williams and Foote have begun the task of discovering whether one such region is really educationally retarded, and if so how much. Periodic sur-

veys, by means of educational tests, of favorable and unfavorable educational environments would help to bring national aid to the schools which most need help. To educational measurement must fall the tremendous task of national diagnosis, and of setting up minimum educational standards for the country as a whole.

CHAPTER VI

MEASUREMENT IN VOCATIONAL GUIDANCE

Functions of Vocational Guidance.—The investigations reported by Davis justify the conclusion of Parsons that we guide our boys and girls to some extent through school, and then drop them into this complex world to sink or swim, and, if they swim, to drift into some line of work by chance, proximity, or uninformed selection.

Vocational guidance, then, has a genuine function, namely, to help each individual reach that particular vocational niche or, better, gateway which leads where he will most greatly benefit himself and most fully contribute to the good of all. This may mean, according to circumstance, an occupation with a fair wage, continual opportunity for self-improvement, and a prospect for advancement; or it may mean a job where the wage is small but which is in line with his ambition and in which he will receive the needed training; or it may mean only a temporary position which will provide the quickest and largest financial return in order that he may be able to resume his school preparation for his chosen field. A realization of this function means that vocational guidance must deal not only with what Clark calls the "misfit human material that has formerly gone into society's scrap-heap," but with the misfit material that never reaches the scrap-heap stage.

Another important function of vocational guidance is to bring into the pupil's education the drive of what President Eliot calls "the life career motive." Some pupils can joyously study, like the monk washed dishes, for "the glory of God," but such pupils are rare. To some no education is

liberal except one which "flutters in all directions and flies in none." There is no reason why an agricultural training should be essentially mean. Some of the most cultural phases of any education center about nature, the farm, and country life. The most beautiful portions of all poetry is the poetry of nature. The testimony is well nigh universal that the life career motive increases studiousness, decreases elimination, and reacts most favorably upon the teachers and supervisors as well.

These functions can be achieved only through (1) a careful survey of the various occupations to determine the constancy of demand for employees, whether the occupation is a seasonal or ephemeral one, the ratio of demand to supply, the monetary rewards, the nature and amount of other types of rewards, the working conditions in the occupation, etc.; (2) a study of the results of such a survey by the pupil, both to aid him to choose his own occupation intelligently and as an important part of his general education; (3) a testing in various ways of the pupil's ability for and interest in each of the occupations; (4) the choice by the pupil with the advice of a vocational counselor, of his vocation; (5) the provision of adequate vocational education; (6) appropriate educational guidance in the light of the chosen vocation; (7) vocational placement at the end of the pupil's educational preparation; and (8) a systematic follow-up of each pupil sent into industry. Only that portion of this total process which is most intimately related to measurement will be discussed further.

Intelligence Limits in Vocational Guidance.—A boy of twelve or a youth of twenty stands before some school official enquiring what occupation it would be advisable for him to enter or for which to begin preparation. What must the educator know before he can give wise advice, and how can measurement help in this intensely human situation?

Sound advice requires the educator or vocational counselor to know the general intelligence limits of the various occupations. This means that intelligence tests must be applied to

members of representative occupations. Terman[1] has made some progress in the determination of occupational intelligence limits. The overlapping of I.Q.'s for the different occupations is so great that some college students have less intelligence than some hoboes! The median I.Q. and more especially the Q_1 more nearly bring out the true facts, namely that success as a business man or college student requires an I.Q. considerably in excess of that which is typical for hoboes, salesgirls, firemen, policemen, motormen, and conductors.

A War Department bulletin on army[2] mental tests shows the intellectual level for various occupations as determined by the application of thousands of intelligence tests at the army cantonments. The scores on these tests, for occupations shown, follows:

45 to 49—Farmer, laborer, general miner and teamster.
50 to 54—Stationary gas engine man, horse hostler, horseshoer, tailor, general boilermaker, and barber.
55 to 59—General carpenter, painter, heavy truck chauffeur, horse trainer, baker, cook, concrete or cement worker, mine drill runner, bricklayer, cobbler, and caterer.
60 to 64—General machinist, lathe hand, general blacksmith, brakeman, locomotive fireman, auto chauffeur, telegraph and telephone lineman, butcher, bridge carpenter, railroad conductor, railroad shop mechanic, locomotive engineer.
65 to 69—Laundryman, plumber, auto repairman, general pipefitter, auto engine mechanic, auto assembler, general mechanic, tool and gauge maker, stock checker, detective and policeman, toolroom expert, ship carpenter, gunsmith, marine engineman, hand riveter, telephone operator.
70 to 74—Truckmaster, farrier and veterinarian.

[1] Lewis M. Terman, *The Intelligence of School Children,* p. 286; Houghton Mifflin Company, N. Y., 1919.
[2] *Army Mental Tests, Methods, Typical Results and Practical Applications;* November 22, 1918, Washington, D. C.

75 to 79—Receiving clerk, shipping clerk, stock-keeper.
80 to 84—General electrician, telegrapher, band musician, concrete construction foreman.
85 to 89—Photographer.
90 to 94—Railroad clerk.
95 to 99—General clerk, filing clerk.
100 to 104—Bookkeeper.
105 to 109—Mechanical engineer.
110 to 114—Mechanical draughtsman.
115 to 119—Stenographer, typist, accountant, civil engineer, Y. M. C. A. secretaries, medical officers.
125 and over—Army chaplains, engineer officers.

The first step is to utilize tests to define the intelligence limits of the various occupations. The second step in vocational guidance is to measure the individual to be guided to determine in which occupation level his intelligence falls. Then the vocational counselor is in a position to tell the pupil the work he is by intelligence fitted to do. The pupil can be informed that his intelligence approximately equals the average of that of individuals who are successfully engaged in, say, ten different occupations. The pupil may, if he chooses, decide for an occupation that is in the next intellectual level above, but he will not do so without being warned that the higher he aims above his natural level the smaller become his chances of success. Good luck, family pull, the possession of valuable accessory traits, etc., may cause him to "get along" out of his intelligence element, but he should realize that the attempt would be a speculative one.

Such a determination of a pupil's intelligence is not only advantageous to the pupil, it may be very profitable for an employer, particularly if the employer has an opportunity to choose among applicants. Recently an almost physically perfect youth was given an intelligence test by a member of our psychology department. The test showed him to be feebleminded. Shortly afterward he was employed as a

messenger boy by Wanamaker. A package entrusted to him disappeared. Detectives watched the boy and annoyed members of his family for several days. Later the package was found in the store where it had been carelessly dropped. At the end of the first week the boy was paid and dismissed. He lost his money before reaching home. Several other employers discovered their mistake by the same trial-and-error expensive procedure. Neither the boy nor his family nor his employer profited by these experiences.

The determination of vocational intelligence limits and the placement of a pupil between these limits presents one great obstacle. We cannot now measure *general* intelligence. General intelligence is, according to Thorndike, composed of three intelligences, namely, abstract intelligence, social intelligence, and mechanical intelligence.

Within any one of these intelligences an individual displays great consistency, but while the correlation between any two intelligences is positive it is not as high as between abilities within any one intelligence. People who have abstract intelligence are able to deal successfully with abstract ideas. They are about equally successful with scientific principles, mathematical and chemical formulæ, legal distinctions and the like. They make capable lawyers, scientists, and theologians. Those endowed with social intelligence are most successful in handling human situations involving human relationships. If they have a social intelligence plus a strong instinct for mastery they tend to make successful business executives, army commanders, and other leaders of men. If the instinct for submission predominates they make satisfactory salesmen, politicians, saleswomen, or *wives*. Those with a mechanical intelligence prefer and are most competent to manage automobiles, motor boats, aëroplanes, washing machines, carpet sweepers and other mechanisms.

To what extent this triple differentiation is caused by original nature, and to what extent by experience is unknown. That most men are stronger in one of these fields

than in the others will be readily admitted. That the types continuously merge into each other and that most individuals who are very successful in one field could have learned to be a fair success in the others also, is not so generally understood. Nevertheless there is a sufficiently genuine distinction to make the differentiation useful.

These three intelligences cannot be measured with equal accuracy and ease. The traditional intelligence tests measure abstract intelligence primarily, though they do to a certain extent measure directly the other two also. Stenquist and others have made some progress toward the construction of tests for measuring mechanical intelligence. The measurement of social intelligence in any satisfactory fashion is the most difficult and consequently least developed of the three.

Moral and Physical Limits in Vocational Guidance.—An individual's vocational fitness depends upon his intellectual, moral, and physical abilities and purposes. Though individuals who are intellectually competent will be found to be on the average morally more dependable, the two are not identical by any means. There is a sufficint absence of correlation to prompt Hollingworth [3] to say:

> "I would rather trust my life and limb to a motorman whose feeble memory span is reënforced by a loyal devotion to the comfort of his grandmother than to a mnemonic prodigy whose chief actuating motive in life is to be a 'good fellow.'"

The correlation between the physical and either mental or moral is, though positive, still less than between the mental and the moral.

There are moral limits in occupations just as surely as there are intellectual limits. For certain types of occupations the moral limits are exceedingly low. The employee in certain low-grade occupations need not be honest, industrious, sympathetic, courteous or much of anything else ex-

[3] H. L. Hollingworth, *Vocational Psychology*, p. 217; D. Appleton & Co., 1916.

cept physically strong. A boss is in constant attendance to see that the employer's property is not stolen and that propensities toward laziness do not injure his interests. We can contrast with this the minimum moral limit required to be a desirable Justice of the Supreme Court, commander of an army in time of war, or any of the numerous positions which carry with them tremendous character responsibilities.

Similarly there are physical limits in occupations. Recent experience has made everyone conscious of the physical limits in a military vocation. The occupations of prize fighting and wrestling each have a high physical minimum. An individual with delicate health and puny physique would scarcely be guided into the occupation of felling timber, heaving coal, digging holes, or breaking stones. Neither would a one-legged youth be guided into a job as messenger boy, or into professional baseball.

The measurement of an individual's physical abilities, in so far as they relate to health, strength and the like is now accurate enough for most practical purposes of vocational guidance. The records of medical inspection and the measurements of physical directors should be carefully preserved for the use of the school's vocational counselor. The measurement of more subtle yet important physical abilities such as beauty, handsomeness and other similar physical qualities cannot yet be done objectively. They must be measured somewhat as moral qualities are measured, so that the two may be considered together below.

Unlike the measurement of intelligence the measurement of an individual's moral abilities and of certain subtle physical qualities must, for the present at least, be entirely subjective. The measurements must be made, either in terms of what associates think of an individual's honesty, courtesy, loyalty, industry, beauty, etc., or in terms of what the individual himself thinks concerning how these qualities of his impress other people. These methods are respectively that of *consensus of associates* and *self-analysis.*

The method of consensus of associates must be accepted as more accurate than the method of self-analysis, particularly if the associates are competent judges, are numerous enough, are sufficiently acquainted with the individual being judged, and have no conscious or unconscious motive for over-rating or under-rating the individual. The difficulty of securing judgments which even approximate these criteria makes it important to determine the accuracy of an individual's rating of himself. Presumably an individual knows himself better than any other individual, and, when seeking vocational guidance, he may have no decided motive for exaggerating his excellencies or minimizing his faults.

Hollingworth reports in *Vocational Psychology* a particularly thorough study of the accuracy of self-analysis.[4] Cattell, Norsworthy, Wells, Fairchild, and others have also contributed to our knowledge of these measurements. Hollingworth selected from about one hundred and fifty women students in their third college year twenty-five students all of whom were acquainted with one another. He asked each student to rate all the twenty-five (the estimator included) in order for their possession of, say, humor, and then in order for conceit and so on for such other traits as neatness, intelligence, beauty, vulgarity, snobbishness, refinement, sociability, kindliness, energy, efficiency, and originality.

His results lead to the following conclusions: (1) The general error of self-estimation tended to be half again as great as the average error of the student's associates. (2) Assuming the average rating of their associates to be correct, there was an average constant error, in the case of self-estimation, toward an under-estimation of their possession of undesirable traits and an over-estimation of their possession of desirable traits. They exaggerated most their possession of refinement, and they minimized most their possession of vulgarity. There was no average constant error for self-estimation of beauty. This of course does not mean

[4] For an outline for self-analysis see, *Outline of a Study of the Self*, by Yerkes and LaRue; Harvard University Press, 1914.

that *each* student was egotistically inclined. Self-estimates from scientific men, secured by Cattell, showed no such constant error. (3) In the case of desirable traits, ability to judge self and others accurately accompanies possession of that quality, whereas in the case of undesirable traits the reverse is true. (4) The student who most accurately rates herself tends to most accurately rate others, though this, like all the conclusions stated above, varies with the trait in question. The reader is referred to Hollingworth's book not only for the details but for additional conclusions.

For practical purposes the best evidence then is that an individual is one-half more inaccurate than the average inaccuracy of his or her associates. We do not know to what extent the error of self-analysis would have been reduced or increased had the students been men instead of women, or had the traits been more carefully defined, or had each student been instructed to rate herself in terms of what others think of her instead of what she thinks of herself, or had the situation been a genuine vocational-guidance situation. The method of self-analysis appears to be sufficiently accurate to have value in vocational guidance.

Aptitude Limits for Vocational Guidance.—The presence in rare individuals of a phenomenal aptitude for some one occupation provoked not only the interest of primitive peoples but gave rise to several systems of magic and incantation by which these aptitudes might be prophesied, and the individual thus guided into a vocation where wealth and honor awaited him. Any one desirous to know the task designed for him by the Fates need only have his horoscope read in the light of his particular star or from the month of his birth. If he doubts the reliability of this horoscope he can have it verified by a palmist who will read the lines in his hand, or a phrenologist who will feel the bumps on his head, or better still a physiognomist who judges anything from the individual's aptitude to the spiritual state of his soul by the shiftiness of his eye, loftiness of his brow, squareness of jaw, thinness of lips, generosity of ear, pro-

trudingness of chin, distribution of dimples, shufflingness of gait, clamminess of hand, and fishiness of eye.

Horoscopes have disappeared except from the advertising columns of frontier newspapers. Palmistry has found its worth level, namely, as an entertaining exercise. Phrenology is still a source of income for its practitioners and is still a potent determiner of social attitudes. Physiognomy is almost universally accepted and is daily practiced.

All four methods, either individually or collectively, have zero or almost zero correlation with what is actually inside the cranium. Physiognomy has a slight but practically negligible correlation with those minute neural connections which are the really significant objects of investigation. But so long as an individual is certain to be judged by phrenological and physiognomical symptoms, phrenology and physiognomy have genuine significance for vocational guidance, for as Hollingworth states: "Vocational success depends not only on the traits one really possesses, but also somewhat on the traits one is believed to possess."

While there are instances of marked special aptitudes for just one occupation, these instances are so rare as to be practically negligible. Each individual probably has an aptitude for some one occupation more than for any other but for most of us it would probably require the exactness of the Infinite to distinguish the occupation. The truth is that most persons, so far as capacity is concerned, could pursue any one of a dozen different occupations with practically equal chances of success. The same qualities which make for success in the medical profession also make for success in the engineering profession or any other profession on approximately the same intellectual level. The idea that most of us have a marked ability for just one of these professions assumes that each profession demands the exercise of only a small proportion of our mental make-up. As a matter of fact each demands the totality of an individual. Each of the higher occupations requires for success an all-

round general ability and about the same all-round general ability.

As a rule there are no round pegs and square pegs. There are big pegs and little pegs. The individual who is always bemoaning the fact that he is a square peg trying to adapt himself to a round hole is probably a little peg trying to fill a big hole. These persons, together with occasional individuals of acknowledged high general ability who have failed, have given currency to the idea that square-pegness and round-pegness is the rule. Truly able individuals frequently fail in one occupation and succeed in another, not because they always possess a special aptitude for the latter only but because the accidents of circumstances chanced to be against them in the former occupation and for them in the latter. Luck is of course far less influential than ability, but its influence is nevertheless considerable. If we think of an individual's different abilities as being the spokes of a wheel, all these spokes tend to be of the same length and the tire of the wheel tends to be a perfect circle. In a few instances there are extreme departures from this circle. In all instances there is probably some departure. The chief difference between individuals is not that one projects in one direction and another in another direction, but that one is a big circle and the other is a little circle.

There is an objectively and practically measurable something which constitutes the core of most aptitudes. It is overlaid with various incidental abilities and furthered or retarded by emotional or physical characteristics of the individual. This something is general intelligence. If an individual's intelligence is all that is known some mistakes will be made in attempting vocational guidance, but if only one thing can be known, general intelligence is perhaps most important, for it is in this that individuals differ most and most significantly. Most men's legs are sturdy enough to carry all the weight of their brain. Most men have enough body to carry them successfully for most occupations.

Most men who possess good intelligence usually have sense enough to realize that they must be fairly honest, decently industrious, etc. Failure is most frequently traceable to lack of brains. A pupil's intelligence score is an approximate measure of the diameter of an approximate general ability circle and is hence an approximate basis for vocational guidance.

But any individual who assumes that all the spokes in an ability wheel are of exactly equal length, or that instances of marked special aptitudes do not exist, or even that most individuals do not possess some tendency toward a special aptitude would make as egregious an error as one who assumed that individuals are all markedly lop-sided. Three principles or near-principles will make clear the limitations of guidance by intelligence tests.

The first principle is that to guide a pupil into a highly specialized occupation requires a specialized series of tests. Certain traits such as mathematical ability, ability in drawing, musical composition, singing, etc., may be so specific as to require a special diagnosis. It is fairly well established that a general intelligence measure will not reveal whether an individual possesses the peculiar combination of traits requisite for success in certain specialized occupations. The *miniature, analogy, analytic,* and *empirical-sampling* types of tests, described in a later chapter, have been used to get at certain of these aptitudes. Thorndike's series of tests for clerical workers, and Seashore's tests of musical capacity, and Rogers' test for the diagnosis of mathematical ability, are all attempts to measure the degree of presence of certain specialized abilities.

The second principle is that the lower we go in the occupational scale and the less the exercise of intelligence is required the less significant is an intelligence measurement as a basis for vocational guidance. Simple computation, checking and the like in clerical work are usually done about as well by persons of moderate intelligence as by persons of high intelligence, for the reason that the exercise of no more

than a rudimentary intelligence is required. And appropriate specialized tests could easily discover individuals of low intelligence who have enough aptitude to actually do better work than individuals of higher intelligence. If one gets down the occupational scale a little farther, a point is soon reached where the aptitude of the horse, dog, or cow surpasses that of man. On the other hand, the higher up the occupational scale one goes, and the more the positions become responsible ones, and the more they require the exercise of a broad general intelligence, the more significant differences in intelligence become for purposes of vocational guidance. A vocational psychologist could serve the selfish purposes of a large industrial concern not only by showing when to choose employees for intelligence, but when there is little or no advantage in so choosing them.

The third possible principle is that disabilities are more frequent than special aptitudes. It is the presence of special disabilities which often explains why an otherwise gifted individual fails to succeed in some highly specialized occupation. Carney [5] describes a typical instance. A graduate of Chicago University, who had an unusually keen mind and a pleasing personality, entered a large factory and was set to work computing percentages on a slide rule. To the surprise of all she failed to do satisfactory work. She was sent to Carney to be tested. She proved to be very high in intelligence and very low in arithmetic. She was assigned to a section which required the continuous exercise of general intelligence. In a short time, she had risen to the head of this section and was doing remarkably well. Thus the intelligence test gave the clue to the existence of a special disability.

Time does not permit the measurement of intelligence, and the measurement of all possible abilities and disabilities. The measurement of the length of all the spokes in any individual's ability wheel is a practical impossibility. It is

[5] Chester S. Carney, "Some Experiments with Mental Tests as an Aid in the Selection and Placement of Clerical Workers in a Large Factory"; *University of Indiana Bulletin,* Vol. V, No. 1, pp. 60-74.

in fact unnecessary. The variation in the length of these spokes from individual to individual is generally so slight, or is so insignificant in terms of vocational success that vocational guidance can ignore them unless the departure from the average is so extreme as to attract immediate notice. Certain traits of an individual need never be measured except in connection with certain occupations. The usual variations in finger length are customarily of no occupational significance, whereas for a typist or pianist their significance may be very great. Variations in beauty are said to be of no significance in the teaching profession, whereas they may have a profound effect upon the success of private secretaries. A cumulative record of a pupil's scores on educational tests given during the school career should be a very great aid and time-saver to the vocational counselor. A pupil whose spelling is abominable is not likely to succeed as a stenographer. A pupil who is slow and inaccurate with numbers will be handicapped as an accountant. A survey is urgently needed to locate the common causes of pronounced failures or pronounced successes in the various occupations in order that vocational counselors may be on the alert for the presence of these traits.

Trade-Ability Limits in Vocational Guidance.—With but very rare exceptions industry expects to employ and then train its employees. These recruits frequently leave or are dismissed before their training period is completed. They seek similar employment elsewhere. The second employer has a different problem from that of the vocational counselor. He must not only measure capacity to learn but must also measure the extent to which the applicant has acquired the specific occupational skills. The vocational counselor will be face to face with an identical problem just as soon as specific vocational training is undertaken on a large scale by the public school system. The problem is in fact already here.

The army psychologists were confronted with the task of this second employer. They met it by analyzing numerous

jobs and by constructing carefully calibrated oral, picture, and performance tests which either measured directly the amount of presence of the occupational skill or measured certain symptoms of occupational skill.

The following sample questions on one of the blacksmith's tests indicate the nature of the oral test. Why is a flatter used? What is shown by white sparks flying from a piece of tool steel when it is in the fire? What do you use for tempering steel springs? What tool is used to make grooves? The picture test presented the candidate with tools, machines, products, etc., of his trade and required him to identify them and indicate their uses. In the expectation that some men could *do* who could not *tell,* some performance tests like truck-driving tests were used. According to Bingham, the expected difference did not materialize except in such rare instances as to be practically negligible and to justify the conclusion "that superiority in trade proficiency resides more often in the head than in the hands." The detailed procedure for the construction of such tests is described in a later chapter.

These trade tests illustrate the technique. They are not nearly adequate for industrial purposes. They are not adequate, first, because they do not reveal how well or how quickly an individual will learn the trade nor whether he has special aptitude for the trade. Intelligence and aptitude tests are required for this purpose. They are not adequate, second, because they are not nearly numerous enough to cover all the many specialized industrial and business processes. Companies, like the Scott Company, are carrying forward, with the coöperation of large industrial companies, the work of adapting the army technique to the construction of needed occupational-ability tests.

Personnel has an amusing description of what resulted in the early days of the late war before trades were defined or detailed trade tests had been constructed. The butchers who went overseas with the first troops were clerks as often as they were butchers. As luck would have it, butchers were

not needed, due to shipments of frozen meat. So the butchers were converted into fairly successful meat bookkeepers. Later troops were accompanied by genuine blood-letters rather than bookkeepers. A bitter complaint regarding their efficiency came from France. Investigation showed that the real need was for "paper butchers instead of meat butchers." Again, a call came for Multiplex puncher operators. Since no one knew that Multiplex *telegraph* puncher operators were a kind of specialized typists, the personnel staff took a chance and filled the requisition with "drill press punchers, ticket punchers, and cow punchers." This is a pretty good description of the vocational placements now effected by chance, where, according to Thorndike, the vocational counselor is a sign which reads: BOY WANTED.

Interest Limits in Vocational Guidance.—The first step in vocational guidance is to determine the mental, moral, physical, special aptitude, and ability requirements of each occupation. The second step is to list for the pupil and his parents the occupations in which his ability promises success. From these a selection may be made on the basis of the pupil's purpose, strength of interest, or preference, because every increase in interest materially increases the chances for occupational success.

What is to be done when interest and intelligence conflict? In this conflict are found some of the real tragedies of life —some striving to be that which they can never be, and some fretting because they have no chance to be that which they could be. In handling the former situation three principles must be kept in mind. *First,* intelligence is comparatively unalterable while interest may be altered either as a result of maturity or artificial manipulation. Interest can be created, intelligence can't. *Second,* success tends to bring interest in an activity which previously was uninteresting. Any activity which brings monetary rewards or the approval of those whose approval is precious tends to become suffused with the pleasure which it purchased. *Third,* after

a certain small interest minimum has been exceeded additional increments of interest are probably less effective in terms of increased success than additional increments of intelligence. This third point is merely a probable hypothesis, for little is known concerning the relative contributions toward success of varying amounts of interest and intelligence either in school or out.

The changeable nature of interest is a major problem in vocational guidance. The staid business man originally intended to be a locomotive engineer. The Greenwich Village poet meant to be an army sergeant.

In so far as the content of elementary and high school subjects is representative of occupational content, a study by Thorndike [6] of the interests of college students leads to the following conclusions: (1) That there is considerable permanence of pupils' interests; (2) that there is an equal permanence of pupils' abilities; (3) that the transition from high school to college marks a more drastic change in interests and abilities than the transition from elementary school to high school; (4) that there is no point where interests and abilities become markedly stabilized; (5) that pupils' interests are highly indicative of their abilities.

Bridges [7] and Dollinger conducted an investigation to check Thorndike's conclusion that pupils' interests are highly indicative of their abilities. All their correlations were very low, around .2 and .3. Consequently they concluded that interest is not highly indicative of ability and that somehow interest and ability must be measured separately. Thorndike would undoubtedly subscribe to the latter half of their conclusion.

For the purposes of vocational guidance Crathorne's [8] investigation is more pertinent. He studied the change of

[6] E. L. Thorndike, "Early Interests: Their Permanence and Relation to Abilities"; *School and Society*, Feb. 10, 1917.

[7] J. W. Bridges and Vernon M. Dollinger, "The Correlation Between Interests and Abilities in College Courses"; *Psychological Review*, July, 1920.

[8] A. R. Crathorne, "Change of Mind Between High School and College as to Life Work"; *School and Society*, Jan. 3, 1920.

mind between high school and college as to life work in the case of 2,000 college freshmen. He found that on entering the high school 57 per cent of these students had decided upon a life work. Before entering college almost exactly 50 per cent of these 57 per cent had changed their minds. And it was the privilege of the men to change their minds as frequently as the women! If it may be assumed that all students who go to high school are in the same mental state with reference to a vocation as those who enter college, it is fair to conclude that of high school freshmen only about 25 per cent have made anything like a stable vocational choice. The mental state of those who do not go to high school is not known. It appears as though a student does not finally make up his mind until he is in a vocation, and even then he frequently changes his choice. Perhaps this uncertainty is desirable, perhaps it is not. Educators must consider whether the school should help to make possible an early and stable vocational choice by giving a wide variety of occupational experiences in the school.

Properly to perform all of the foregoing tasks requires a vocational counselor who possesses not only rare insight but also technical training[9] of a very high order. He must know how to analyze occupations to discover the traits required for success and in what proportion they are required. He must be able to construct tests which will measure these traits. He must be able to select tests which correlate with demonstrated fitness for a given occupation. He must know how to select the proper team of tests weighting each test in the team according to (a) the largeness of its self-correlation, (b) the smallness of its correlation with the other tests in the team, and (c) the largeness of its correlation with the criterion. He must know how to apply these tests and how to select the pupils who can learn an occupation and be

[9] For an admirable brief summary see Truman L. Kelley, "Principles Underlying the Classification of Men," *The Journal of Applied Psychology*, March, 1919, and E. L. Thorndike, "Fundamental Theorems in Judging Men," *The Journal of Applied Psychology*, March, 1918.

efficient and happy in it. Ultimately he will have the additional problem of distributing the available pupils into the available occupations so as to secure the minimum of misplacement.

Vocational Guidance for the Gifted Pupils.—A great social waste is the vocational exploitation of the unusually gifted. With certain exceptions every employer is competing with other employers to secure the services of the most competent. The employer does not stop to consider whether he can give the gifted individual, whom he is lucky to employ, abundant opportunity to make the greatest social contribution of which he is capable. The country suffers an enormous loss each year because many of its geniuses have been caught by this exploiting system, and placed in relatively non-productive positions. The individual employer can afford this but society can't. Society's aim is to guide no individual into an occupation above his intelligence. Society is equally concerned that great gifts be not frittered away on small jobs. In sum, we want both minimum and maximum intelligence limits for each occupational level. In so far as it can be done without doing too much violence to individual liberty, the social group should guide each individual to the level fixed for him by nature. Only thus can the social group be most efficient, prosperous, and happy.

In time society will recognize its essential organic nature, and then the persons of low and average ability will themselves insist that the able be placed where they can make the greatest contribution for the good of all. The gifted, considering their superior native endowment as part payment for their services, will contribute to the social group without extorting undue monetary rewards from the group which they serve. Vocational guidance through the schools is about the only way to accomplish this great and beneficent task.

Society cannot safely trust its geniuses to find their own way through the industrial maze. Immature occupational

preferences frequently lead where there is no turning back. Below are selections from a list prepared by Miss Coy [10] and another by Mrs. MacKnight of the vocational ambitions of gifted children.

1. "I would like to be an author as it has a certain fascination and I have a rather flaming imagination. Also, you get a certain amount of royalty on each copy of a book that the publisher accepts. This would keep any successful writer well provided for." (Age = 11. I.Q. = 143.)

2. "When I will be a man, I will be in the business that my father is in. That business is the Stock Exchange. The reason is that you don't sit down like most of men who work." (Age = 8. I.Q. = 143.)

3. "When I grow up I would like to be a research scientist of electricity, because I like to work with electricity." (Age = 9. I.Q. = 145.)

4. "When I grow up I think I will be a man who runs a boat a with mail up and down the lake and earn my living that way." (Age = 8. I.Q. = 151.)

5. "When I grow up I would like to be an inventor. The reason why is because I told my father what I was going to invent and he said it was all right." (Age = 9. I.Q. = 126.)

6. "I would like to be a surgeon because I would help people." (Age = 10. I.Q. = 131.)

7. A stenographer or music teacher. (Age = 10. I.Q. = 141.)

8. Carpenter or mechanic. (Age = 11. I.Q. = 121.)

9. A soldier—"not a general or a hero, but just a common soldier." (Age = 12. I.Q. = 133.)

10. League baseball pitcher, motorcycle racer, pole vaulter, wrestler, and be an "honest man." (Age = 12. I.Q. = 122.)

Miss Coy concludes from her study of the vocational preferences of gifted children that in the main few of the

[10] Guy M. Whipple, *Classes for Gifted Children*, p. 77; Public School Publishing Co., 1919.

pupils want to do things for which they lack ability but that there is a tendency to report ambitions that seem distinctly too low. She further concludes that efforts to properly educate superior pupils should include a systematic effort "to foster and develop ambitions commensurate with the latent capacities revealed by objective testing."

Vocational Guidance for the Intellectually Subnormal.—Some fear that the guidance of pupils of low ability into occupations with low intelligence minima is essentially undemocratic. Perhaps it is, according to some definitions of democracy. What of it, if such guidance is the best thing to do? So long as occupational levels are not closed to those in a lower level, such guidance is a genuine kindness to the individual and decidedly advantageous to the social group as a whole.

It is a genuine kindness to the individual for at least three reasons. In the first place, he will succeed better in occupations requiring little intelligence. He will succeed not so much because his ability is greatest for these occupations as because here and only here will he be in competition with his own kind. Most of the abler individuals will have risen to occupational levels where more of the world's rewards are for distribution.

In the second place he will succeed better because the public is willing to pay him for services rendered on this level. There are men making a reasonably good living as street cleaners who would starve as school teachers. The public is willing to pay them for cleaning the streets but is utterly unwilling to pay them a cent for teaching school, or practicing medicine.

Finally, the mentally subnormal are probably happier at work which is physical and routine. There are numerous persons to whom, like Javert, thought is singularly painful. Consider the per cent of individuals who would beg to dig ditches or pound rocks to avoid writing this book or even reading it.

SUPPLEMENTARY READING FOR PART I

Ballard, Philip B.—*Mental Tests;* Hodder and Stoughton, Ltd., Warwick Square, London, 1920.

Buckner, Chester A.—*Educational Diagnosis of Individual Pupils;* Teachers College, Columbia University, New York, 1919.

Bloomfield, Meyer.—*Youth, School, and Vocation;* Houghton Mifflin Company, New York, 1915.

Burgess, May Ayres.—*Measurement of Silent Reading;* Russell Sage Foundation, New York, 1920.

Chapman, J. Crosby.—*Trade Tests;* Henry Holt & Company, New York, 1921.

Chapman, J. C., and Rush, G. P.—*Scientific Measurement of Classroom Products;* Silver, Burdett & Company, Boston, 1917.

Courtis, S. A.—*The Gary Public Schools: Measurement of Classroom Products;* General Educational Board, New York, 1919.

Davis, Jesse B.—*Vocational and Moral Guidance;* Ginn & Company, New York, 1914.

Dewey, Evelyn; Child, Emily; and Ruml, Beardsley.—*Methods and Results of Testing School Children;* E. P. Dutton & Company, New York, 1920.

Fretwell, Elbert K.—*A Study in Educational Prognosis;* Bureau of Publication, Teachers College, Columbia University, New York, 1919.

Hollingworth, H. L. and L. S.—*Vocational Psychology;* D. Appleton & Company, New York, 1916.

Hollingworth, L. S., and Winford, C. A.—*The Psychology of Special Disability in Spelling;* Teachers College, Columbia University, New York, 1918.

Huey, Edmund B.—*Psychology and Pedagogy of Reading;* Macmillan Company, New York, 1908.

Various Conferences on Educational Measurement; Indiana University Bulletins, University of Indiana, Bloomington, Indiana.

JUDD, CHARLES H.—*Measuring the Work of the Public Schools;* Cleveland Education Survey, Russell Sage Foundation, New York, 1916.

JUDD, CHARLES H., and OTHERS.—*Reading: Its Nature and Development;* University of Chicago, Chicago, 1918.

KELLEY, T. L.—*Educational Guidance: An Experimental Study in the Analysis and Prediction of Ability of High School Pupils;* Teachers College, Columbia University, New York, 1914.

KRUSE, PAUL.—*The Overlapping of Attainments in Certain Grades;* Bureau of Publication, Teachers College, Columbia University, New York, 1918.

LINK, H. C.—*Employment Psychology;* Macmillan Company, New York, 1919.

MONROE, WALTER S.—*Measuring the Results of Teaching;* Houghton Mifflin Company, New York, 1918.

MONROE, WALTER S.; DE VOSS, J. C.; and KELLY, F. J.—*Educational Tests and Measurements;* Houghton Mifflin Company, New York, 1917.

MUNSTERBERG, HUGO.—*Psychology and Industrial Efficiency;* Houghton Mifflin Company, New York, 1913.

NATIONAL SOCIETY FOR THE STUDY OF EDUCATION.—Various *Year Books;* Public School Publishing Company, Bloomington, Illinois.

PINTNER, RUDOLF, and PATERSON, DONALD.—*A Scale of Performance Tests;* Warwick & York, Baltimore, 1917.

ROGERS, AGNES L.—*Experimental Tests of Mathematical Ability and Their Prognostic Value;* Bureau of Publication, Teachers College, Columbia University, New York, 1918.

STARCH, DANIEL.—*Educational Measurements;* The Macmillan Company, New York, 1917.

STRAYER, GEORGE D., and OTHERS.—*Report of a Survey of the School System of St. Paul, Minnesota;* Department of Public Instruction, St. Paul, Minnesota, 1917.

TERMAN, LEWIS M.—*The Measurement of Intelligence;* Houghton Mifflin Company, New York, 1916.

TERMAN, LEWIS M.—*The Intelligence of School Children;* Houghton Mifflin Company, New York, 1919.

THEISEN, W.—*Report on the Use of Some Standard Tests for 1916-17;* Wisconsin State Department of Public Instruction, Madison, 1918.

WHIPPLE, GUY M.—*Classes for Gifted Children;* Public School Publishing Company, Bloomington, Illinois, 1919.

WILSON, G. M., and KREMER, J. HOKE—*How to Measure;* The Macmillan Company, New York, 1921.

PART TWO

HOW TO CONSTRUCT AND STANDARDIZE TESTS

CHAPTER VII

PREPARATION AND VALIDATION OF TEST MATERIAL

I. Logical and Experimental Validation of Tests

What Is a Valid Test?— Tests have many characteristics such as validity, reliability, objectivity, etc. Of all these traits validity is most fundamental. What is meant by validity? The National Association of Directors of Educational Research has defined validity as the correspondence between the ability measured by the test and ability as otherwise objectively defined and measured. When a test really measures what it purports to measure and consistently measures this same something throughout the entire range of the test it is a valid test.

Ask a cautious psychologist just what a given test measures and he will answer somewhat viz: "It measures the ability to do so and so with the material which you see on the test sheet, when the test is applied under certain conditions." If you are dissatisfied with this conservative statement you may enquire: "Will the pupil who deals with these test difficulties with a given degree of excellence deal with these apparently same difficulties when imbedded in a real, practical life situation with an equal degree of excellence?"

No one knows very much about just how close results for the different tests are to the results in actual practice. We give a class a paper test composed of twenty reasoning problems in arithmetic. Johnny does eighteen of the twenty problems. Had he met these twenty problems at the store or the post office or the playground, would he have succeeded

with these same eighteen problems and failed on the same two? Nobody knows. If he did not do the eighteen but did do sixteen, would Mary and Lucy who did fourteen and twelve test problems respectively show proportional decreases when faced with real problems or might they possibly surpass Johnny? In other words, if test results and life results do not coincide, do they even correlate, i. e., does the pupil who makes the highest test score make the highest life score and the one who makes the second highest test score make the second best life score and so on? Nobody knows. We know enough to say that there will be a rough correspondence and probably a close correspondence, for the chasm between test conditions and life conditions does not yawn as wide as some would have us believe. It is undoubtedly wider for some tests than for others.

The layman is usually concerned with this problem of the correspondence between tests results and practical life results. The technical worker in measurement is equally concerned to know whether a test is a valid measure of some element of an analyzed ability. The following suggestions as to how to secure validity apply primarily to securing a close correspondence between test results and practical life results.

Suggestions for Increasing Validity.—Test results are more comparable to life results the more nearly the test process approaches the character of the life process. The ability of pupils to spell, for example, may be determined by (a) searching through their letters, compositions and the like, (b) having them write dictated sentences in which the critical words are imbedded, (c) having them write isolated words pronounced by the examiner. The composition method more nearly duplicates the life process, the dictation method next, and the column spelling last. Again, pupils' ability in grammar can be measured by making an analysis of their written or oral compositions or by giving them a specially devised grammar test. The former test would yield more natural results. It is of course one of the per-

versities of fate that an increase in naturalness is attended by an increase in inconvenience.

Tests vary greatly in the exactness with which they reproduce the life process. Hollingworth [1] lists four fundamental types of tests: miniature, sampling, analogy, and empirical.

He writes that in the case of the *miniature* test the "entire work, or some selected and important part of it, is reproduced on a small scale by using toy apparatus or in some such way duplicating the actual situation which the worker faces when engaged at his task. Thus McComas, in testing telephone operators, constructed a miniature switchboard and put the operators through actual calls and responses, meanwhile measuring their speed and accuracy by means of chronometric attachments."

The *sampling* test measures a candidate's ability to do an actual sample instead of a toy representation of a given occupation. A would-be stenographer is given an actual test of ability with dictation and with a typewriter. A clerical aspirant is set to finding addresses in a telephone directory or copying a table of figures. Practically all educational tests are dummy samplings of this variety. We test a pupil's reading ability by samples of reading material. We test his ability to solve problems in arithmetic by giving him sample problems in arithmetic to do.

The *analogy* test employs material which is neither the same as nor similar to the material of the occupation, but it is supposed to exercise those mental traits requisite for success in the occupation. To quote Hollingworth again: "Thus girls employed in sorting steel ball-bearings, and also typesetters, have been selected on the basis of their speed of reaction to a sound stimulus." During the World War, Stratton, Henmon, Thorndike and others attempted to devise tests which would be diagnostic of ability for flying. At that time no empirical tests existed, and dummy tests were impractical. So those who were working on the problem first made an analysis of the mental and physical character-

[1] H. L. and L. S. Hollingworth, *Vocational Psychology;* D. Appleton & Co.

istics upon which success in flying would logically seem to depend, and then devised means for measuring a candidate's possession of these traits. Tests were devised to measure a candidate's sense of balance, perception of tilt, nerve-resistance to sudden sensory shock and the like. By checking each of these tests against subsequent success as aviators, it was found that some had no significance at all, while others were slightly symptomatic. A composite score from those tests which were found valuable, selected aviators with fair accuracy. In similar manner tests were constructed to select shell inspectors, gun assemblers, etc. Pursuing this same method of analysis, Rogers has constructed tests for determining whether pupils possess mathematical capacity. Briggs has constructed similar tests for foreign language capacity.

The *empirical* tests are those which were discovered from a more or less haphazard trial and error search. The test selector makes no conscious attempt to select or construct a test which is a miniature or sampling or analogy. He tries out a number of tests, eliminates those which are not symptomatic and retains those which are. Analysis of the mental traits and their combinations requisite for success in the various educational or industrial occupations has not yet progressed far enough to offer a sure basis for the selection and construction of tests. Hence it is the opinion of many psychologists that the empirical method of discovering occupational tests is more promising than any other for the immediate future.

Test results are more comparable to life results when they are free from irrelevancies. To return to the illustration of a reasoning test in arithmetic, the arithmetic problems probably more nearly duplicate real problems when they are free from non-arithmetical difficulties. Complicated instructions for the test might so confuse the pupils as to leave no fair opportunity to attack the arithmetical difficulties. Again, a complex wording of the problems might make the linguistic difficulty of greater consequence than the difficulty

of the mathematical processes themselves. In selecting or constructing tests they should be carefully studied to discover whether everything possible has been done toward the elimination of irrelevancies in instructions and in the organization and wording of the test elements, or at least toward determining the influence of these irrelevancies.

While linguistic irrelevancies are more common, they are not the only kind by any means. The form of the test is often an irrelevancy. Not only must the pupil overcome the difficulties of the real test material, which is always to some extent camouflaged by linguistic irrelevancies, but he must also overcome the difficulty of the general form in which the test is couched. These moulds for test material are many. There are the *question* mould, *completion* mould, *classification* mould, *matching* mould, and many others. All these irrelevancies are important elements of difficulty especially for young children. They do greatest harm in rate tests where the speed score of the pupil is much influenced by the rapidity with which he adapts himself to the test.

Terman [2] says of the army intelligence test Alpha: "The test questions were ingeniously arranged so that practically all could be answered without writing, by merely drawing a line, crossing out or checking." There were various reasons for this provision, such as to require less time for testing and to make scoring economical and objective. But a very important reason was to make a test which would test the thing for which the test was designed. It was designed to measure general intelligence. If writing were made a prominent feature of the test, the test would tend to give a measure of speed of handwriting rather than of intellectual ability. Individuals are more alike in their speed of checking, crossing out and underlining than they are in speed of penmanship.

It is possible, especially in the case of very long tests, that the chief factor measured is not the ability desired but

[2] *Psychological Bulletin*, June, 1918.

fatiguability. The test should be of such a length or so constructed as to eliminate fatigue, particularly if some of the pupils fatigue more easily than others. This point needs most attention when comparisons are to be made between young and old children.

Fatigue may be eliminated in various ways. First, the test can be made short. Second, if reliability requires a longer test, the test can be divided into parts with a rest or exercise interval between. Third, if the test consists of a series of short tests, the shorter tests may be so arranged as to have difficult tests followed by easy tests and tests of one nature followed by tests of another nature and vice versa. Fourth, the test can be made variegated and interesting both as to type and material. The material in the Alpha intelligence test for the army, for example, kept the recruits in a merry and at times almost boisterous mood throughout.

The foregoing propositions concerning irrelevancies should be accepted with caution and applied with care. The propositions were made more to direct attention to certain problems rather than because they have a firm experimental basis. If the examiner's purpose is to make a psychological study of pure arithmetical abilities there can be no question but that every possible linguistic or other irrelevancy should be eliminated from the tests used. Similarly when linguistic ability is being measured, all non-linguistic difficulties should be eliminated. But if life's arithmetic problems are to be duplicated we cannot be so sure of the value of eliminating all irrelevant difficulties. When a child pays for purchases in a store he must steer his course through numerous distractions which are not all mathematical in their nature. Since these practical distractions cannot conveniently be duplicated in a test, perhaps the linguistic or other difficulties should be retained as a sort of substitute. Again, the propositions should be applied with care, because an irrelevancy in one test may not be so at all in another test. If the form or mould of a test duplicates the pattern

of the pupil's mental processes in performing an actual task, the form of the test is not an irrelevancy. A casual inspection of the following task taken from the Woodworth-Wells Directions Test would give one the impression that the whole test is nothing but an irrelevancy, and yet this impression would be a mistake, for the purpose of the test is to measure the ability to deal with just such complicated directions.

"With your pencil make a dot over any one of these letters **F G H I J**, and a comma after the longest of these three words: boy mother girl Then, if Christmas comes in March, make a cross right here........, but if not, pass along to the next question, and tell where the sun rises............ If you believe that Edison discovered America, cross out what you just wrote, but if it was someone else, put in a number to complete this sentence: 'A horse has .... feet.' Write *yes,* no matter whether China is in Africa or not ....; and then give a wrong answer to this question: 'How many days are there in the week?' .... Write any letter except *g* just after this comma, and then write *no* if 2 times 5 are 10 .... Now, if Tuesday comes after Monday, make two crosses here ....; but if not, make a circle here .... or else a square here .... Be sure to make three crosses between these two names of boys: George .... Henry. Notice these two numbers: 3. 5. If iron is heavier than water, write the larger number here ...., but if iron is lighter write the smaller number here .... Show by a cross when the nights are longer: in summer? .... in winter? .... Give the correct answer to this question: 'Does water run uphill?' .... and repeat your answer here .... Do nothing here (5 + 7 = ...), unless you skipped the preceding question; but write the first letter of your first name and the last letter of your last name at the end of this line:"

Increasing Validity Through Comprehensive Measurement.—A test can be made comprehensive, and in a certain sense more valid, by including all the material within the subject or ability being measured. This is feasible when the examiner is interested in only a narrow ability or limited field of subject matter. It is possible to comprehensively measure even a comprehensive subject matter but to do so might require almost as much time as it took the pupil to learn it. Hence some more economical method must be found for measuring a comprehensive ability.

A test can be made comprehensive by including random samplings of the ability in question. In order to determine how many words a pupil can spell, or define, or use, it is not necessary to try him on every word in Webster's Dic-

tionary. It can be done just as well by taking from the dictionary a random sampling of its words. In making such a sampling it is important that the samplings be made random, and that enough samples be employed to yield a reliable measure of the pupil. Randomness may be secured by using the first or ninth or any other numbered word on each page or each third page or each twenty-fifth page or the like of the dictionary. This will suggest how chance samplings may be made from a variety of subject matter. It is worth pointing out that when test material is selected according to this random-sampling method, the construction of duplicate tests becomes a very simple matter. The value of such duplicate tests will appear later. It should be remembered that the method of random sampling answers only the question: What per cent of a total field of knowledge does a pupil know? Except for the elements in the test, such a test leaves us in ignorance as to just what elements in the field of knowledge the pupil knows.

To overcome this last obstacle, especially in the field of skill tests, it has been suggested that comprehensiveness be secured by using type material. This type principle of selection assumes that each subject involving skill contains typical units or typical processes, and that the pupil's ability in the entire subject is substantially determined by measuring his ability in the type processes. The fundamentals of arithmetic, for example, are supposed to contain certain type processes. The ability to *carry* in addition is one such type process. The ability to fix the decimal point in division is another type process and so on. It is held that a test to be representative of the fundamentals of arithmetic must contain every type process. Ballou and Monroe have each analyzed out these processes in the fundamentals of arithmetic and have constructed tests to measure them.

Monroe [3] has criticized the Woody Arithmetic Scales because Woody did not select examples for his tests primarily

[3] Walter S. Monroe, "An Experimental and Analytical Study of Woody's Arithmetic Scales"; *School and Society*, Oct. 6, 1917.

on a type basis. Monroe contends that Woody sacrificed diagnostic ability to statistical beauty, since Woody retained examples in his scales primarily because of their statistical behavior—because of their difficulty.

Another principle for selecting test material which has come into common use is the social-worth principle. This principle makes comprehensiveness subordinate to relative value. The social-worth principle assumes that the most valuable information for the school will come from testing the pupil's ability to spell only those words, or solve only those problems, or demonstrate a knowledge of only those historical facts which are of greatest social value. The best illustration of a test whose construction has been guided by this principle is the Ayres or Ashbaugh Spelling Test. The Ayres test contains 1000 words which were selected by exhaustive investigations to discover which words were most frequently used. Similar surveys for other subjects will make it possible to construct other tests in accordance with this principle.

Comprehensiveness requires that we not only measure how much a pupil can do and how well he can do it, but also we must measure how rapidly he can do it. This proposition needs no justification, for the practical importance of such a diagnosis of the pupil's habit of work is obvious. At least one major aim of the school is to prepare the pupil for effective participation in the social group. The social group does not want the pupil's ability, nor does the pupil derive much joy or profit from his ability, if he falls below a minimum of speed. Thus the three main dimensions of a pupil's ability are (a) how much or how difficult, (b) how well or how accurately or with what quality, and (c) how rapidly. If reading is to be measured, a test or tests (for frequently all three dimensions cannot well be measured in a single test) should be constructed which will measure all three aspects of reading.

Validation of a Trade Test.—How may we know whether a given test measures the ability which we desire

measured? We know what a test measures only by its correlations. Does a pupil's score on an intelligence test coincide with the school's and world's estimate of this pupil? Does the arithmetic test indicate how well a pupil will be able to work examples or solve arithmetical problems in the store or in those realms for which the school is preparing the pupil?

Two ways of determining this correspondence are available. One method is to give a test to a group of pupils, to preserve the records, to follow up the testing with prolonged careful observation of how well these pupils do in real situations which may or may not be arranged by the investigator, to rank the pupils in order of their ability first, on the test and second, in the real situation, and finally, by the method of correlation described later to determine the correspondence between these two rankings. If the agreement is close the test does measure real ability in the sense that it can rank a group of pupils in order for their possession of the ability in question. An even more careful technique is required to determine the extent to which a pupil will make the identical score in both the test and the life situation.

The second method available is to apply the test which is being validated to a group of individuals whose real ability is already known. If the test distinguishes the different degrees of known merit, we can call the test satisfactory. Ruml, Robinson, Chapman, Meine, Kruse, Wylie, Toops, and others constructed about 100 Trade Tests for the army during the war. As the following quotation from the *Psychological Bulletin*, June, 1918, will show, they employed this second method to determine whether their tests really measured the trade skills which the tests purported to measure. A long quotation is given below in order to show details in the process of trade test construction. Few educational tests are constructed with such careful attention to what the tests really measure. The test is usually assumed to measure what it looks as though it measures.

"When the problem of formulating tests was analyzed, it was seen that certain requirements were fundamental. A good trade test: (1) Must differentiate between the various grades of skill; (2) Must produce uniform results in various places and in the hands of individuals of widely different characteristics; (3) Must consume the least amount of time and energy consistent with satisfactory results.

"While there are all degrees of trade ability among the members of any trade, it is convenient to classify them in a few main groups. Ordinarily the terms Novice, Apprentice, Journeyman and Journeyman Expert (or Expert) are employed. The Novice is a man who has no trade ability whatever, or at least none that could not be paralleled by practically any intelligent man. The Apprentice has acquired some of the elements of the trade but is not sufficiently skilled to be entrusted with any important task. The Journeyman is qualified to perform almost any work done by members of the trade. The Expert can perform quickly and with superior skill any work done by men in the trade.

"It is sometimes desirable that the Trade Test should differentiate between the skill of different members of the same group; for instance, the journeyman group. It is essential that it should differentiate between the journeyman and the apprentice, and the apprentice and the novice. Trade tests devised to make this classification are of three kinds: oral, picture and performance.

"The oral tests are most generally used because they are of low cost and they may be applied to a large number of men in a comparatively short time and without much equipment. They are satisfactory in determining the presence or absence of trade ability and in many instances determine the degree of ability with such accuracy that no other tests are required.

"An Oral Test is developed by passage through twelve stages: (1) Priority, (2) assignment, (3) inquiry, (4) collection, (5) compilation, (6) preliminary sampling, (7) revision, (8) formulation, (9) final sampling, (10) evaluation, (11) calibration, (12) editing.

"*Collecting the Trade Information.*—From time to time

the Personnel Organization of the Army submits to the Central Trade Test Office (Newark, N. J.) a list of trades which are required in Army use and for which tests are urgently needed. Upon the basis of this list, assignments are made to the field staff.

"The field staff then makes thorough inquiry into the conditicns of the trade. Their purpose is three-fold:

"1. To determine the feasibility of a test in this trade. It was found, for example, that the trade of gunsmith was not a recognized trade, though there were gun repairers.

"2. To determine the elements which require and permit of testing. In other words, can men be graded in it according to degrees of skill? In some trades it was found that the trade required simply the performance of a single set of operations and there were no gradations among the members of the trade.

"3. To determine the kinds of tests that can be used. Some trades, such as truck driving and typewriting, are mainly matters of skill and for them performance tests are better than oral tests. Other trades, such as interior wiring and power plant operation, are mainly matters of knowledge. For these trades oral and picture tests are best.

"After having discovered by inquiry that the trade is a recognized trade and can be tested, the field staff proceeds to collect all the information necessary from all available sources; for example, experts of the trade, trade union officials, literature of the trade, trade school authorities, employers and the like. They discover by this means what are the elements of the trade and what constitutes proficiency in it.

"*Compiling the Questions.*—As a result of this collection of information they compile a number of questions, usually forty to sixty, each of which calls for an answer that shows knowledge of the trade. Experience in the formulation of such questions has shown that a good question meets the following requirements:

"1. It must be in the language of the trade.

"2. It must be a unit, complete in itself and requiring no explanation.

"3. It must not be a chance question which could be answered by a good guess.

"4. It must be as short as possible and must be capable of being answered by a very short answer.

"5. It must not be ambiguous; the meaning must be unmistakable.

"After the large number of questions originally formulated has been sifted down by application of these requirements they are used in a preliminary sampling on a number of tradesmen whose answers indicate the merits of the different questions and their grades from easy to difficult. In this sampling, tradesmen from different shops or plants are tried, in order to guard against specialized methods or modes of expression confined to a single locality. At least two examiners work on each set of questions at this stage to get the benefit of more than one point of view for revision.

"This preliminary sampling affords a means of checking on the following points:

"1. Is the test applicable to trade conditions?

"2. Does the test represent good trade practice?

"3. In what way can parts be profitably modified, supplemented or eliminated?

"4. Does the test represent the whole range of the trade from the novice to the expert?

"5. Is it a representative sampling of the whole range of trade processes?

"In the light of the answers to these questions, the test is revised and is then ready to be formulated.

"*Final Sampling.*—Final sampling is made by testing twenty men who are known to be typical representatives of each group (novice, apprentice, journeyman, expert). Among the novices tested are some highly intelligent and mature men of good general knowledge but no trade ability. Examinations are given to men whose record in the trade is already known and who are tested as nearly as possible in the same manner as men in the camps.

"The results of this final sampling are now turned over to the Statistical Department of the Central Trade Test Office. The experts in this department make a careful study

of the results and of the answers to each question. This enables them to determine the relative value of each individual question and the selection that makes a proper balance.

"*Evaluating the Test.*—If a trade test is good, a known expert, when tested, is able to answer all, or nearly all, the questions correctly; a journeyman is able to answer the majority; an apprentice a smaller part, and a novice practically none. This does not mean that each question should be answered correctly by all the experts, a majority of the journeymen, some apprentices but no novices. There are few questions which show this result.

"Other types of questions, however, are more common. Some show a distinct line of cleavage between the novice and the apprentice. Novices fail, but apprentices, journeymen and experts alike answer correctly. There are likewise questions that are answered correctly by nearly all journeymen and experts but only a few apprentices, and questions that only an expert can answer correctly. Each type of questions has its value in a good test. The main requirement is that the tendency of the curve should be upward; a question which is answered correctly by more journeymen than experts or more apprentices than journeymen is undesirable and is at once discarded. A proper balance is made of the others.

"*Calibrating the Test.*—One task still remains; namely, that of calibrating the test. As each question is allowed four points, it becomes necessary to determine how many points should indicate an expert, how many a journeyman, etc. Obviously the way to do this is to note how many points were scored by the known experts and the known journeymen when they were tested. Ordinarily the expert scores higher than the journeyman and the journeyman higher than the apprentice. It frequently happens that a few journeymen score as high as the lowest of the experts and a few apprentices as high as the lowest of the journeymen. There are consequently certain overlappings between the classes. In calibrating, the object is to draw the dividing line between classes so that the overlapping shall be as small as possible.

"When these dividing lines, or *critical scores* as they are usually called, are established, the test is ready for distribution to camps."

Suppose that we give a group of pupils a test in arithmetical problems, and then, without arousing the suspicion of the pupils, arrange the situation so that these same pupils will meet these same arithmetical problems in their play life on the street, and suppose that the test and the observations upon the pupils' success with the play problems are reliable measures of each of these abilities and suppose, finally, that the correlation between the test and the observations is of only average closeness, does this condemn the test as not being a measure of real ability? Assuming that proper experimental precautions have been taken, this correlation certainly tells us that the test problems are a rough but not an accurate measure of play problems. But before we condemn the test we ought to correlate the pupils' scores on play problems with their scores on those same problems when shopping for their mothers or some other practical situation. It is not known, but it is very possible that the correlation between different real-life situations is no closer than between the test and any one of these situations. In sum, it is even probable that there is no such thing as real ability, in the sense that we are discussing it, but that there are instead, many abilities differing somewhat one from another. It is hopeless to expect to find a test which will closely correlate with each of these life situations, wrapped about, as each is, with its own individuality or specificness.

It might be possible to eliminate experimentally, all of the specificness belonging to our test and each life situation, and thus demonstrate a perfect correlation between all the thus purified abilities. Such an analysis of abilities would be of considerable theoretical interest and, as we saw in our discussion of diagnosis, would be of great value in connection with remedial instruction. But for the purpose of prophesying success in life and the like, we cannot deal with these rarefied abstractions of abilities, for abilities must always

function through specific situations. It would have been of no comfort to Ruml and his colleagues to know that their Trade Tests, when experimentally purified, correlated perfectly with similarly purified trade situations. They were asked to construct tests which would, with the least error, select men who could make good in a variety of specific situations. It is no condemnation of an educational test if it shows only substantial correlation with a variety of real situations. It is a condemnation when it shows little or no correlation with real abilities or when it shows less correlation with such abilities than some other available test which is equal in all other requirements.

II. Validation of an Intelligence Test

Criterion of Intelligence.— How may we determine whether a particular test yields a valid measure of intelligence? As was stated a few pages back, *what a test really measures is known only by its correlations.* The closeness of a test's correlation with what constitutes intelligence is a measure of its excellence as an intelligence test. This is the single ultimate method of choosing such a test. If this condition is satisfied the examiner need look no further. In order to determine a test's worth, it is necessary to have a number of pupils accurately rated for intelligence. Tests can then be correlated with this rating.

This intelligence rating is usually secured by having the most competent obtainable persons observe the objective behavior of the pupils and, allowing for non-intelligence factors, estimate the degree of intelligence indicated by this objective behavior. There is no reason why these observations should not extend over a life time and then be checked by historical verdict. By means of a rating so obtained, we can, by the trial-and-error process, find a test which gives results which closely agree with the observations of competent observers. Once one test has been carefully validated in this fashion, other tests can be constructed and correlated with the original test.

If social judgment is the final criterion of intelligence, why not employ it exclusively? Tests are resorted to not only because they are far more economical but particularly because they are impersonal and prophetic. History has changed too many pillories to monuments, and parental evaluations of children have been too frequently reversed for us not to know that subjective judgment often tends to be prejudiced unless the observations are safeguarded with unusual care. For this reason the relatively ice-cold mob-proof, carefully-validated intelligence tests are coming to be used more and more for intellectual determination. Intelligence tests possess the further incalculable value of being prophetic. Tests do not wait until success is achieved before passing verdict. They do not lay roses on the grave of a genius, but crown him in childhood.

Must intelligence tests always be searched for in the exceedingly wasteful trial-and-error fashion? Cannot the psychological analysis of intelligence be carried far enough to at least indicate the general direction for test construction? It can, but psychological analysis is no substitute for experimental evidence of validity. For illustrative purposes such a tentative analysis, in terms of the neural mechanism, is outlined below.

Analysis of Intelligence.— 1. *The number of desirable neural connections.*—One of psychology's important criteria of superior or inferior intelligence is the differences in the ability for minute analysis, "piece-meal activity," or to deal with subtle elements of a situation. In neural terms this means that the more intelligent individual has more neural connections for any one situation. To a stupid individual a ripe peach will probably suggest gastronomic satisfaction only, while to the more intelligent it suggests this to be sure, but it may also suggest the flush of dawn, the blush of a maid, the softness of a baby's cheek, or the fruit of the Tree of Knowledge! The flower in the crannied wall was more to Tennyson than a pretty weed to adorn a vain buttonhole! To those with numerous neural connections "every chip

sprouts wings to bear a god" and falling apples cause a flow of ideas as well as a flow of saliva! To a Woodberry, "a rose shadows us with Persia, or a single lotus blossom unbosoms all the Nile."

2. *The system of organization of connections.*—Mental "short cuts" and "hierarchies" are essential for effectiveness in novel thinking and highly skilled action. And a hierarchy is simply a very complex coördination of neural connections. System, coördination, organization are as important in the mental sphere as they are in a telephone system or in any social, business, or industrial sphere. While educational experts disagree as to whether the unit of organization should be projects or general principles or something else, all agree that education should result in neural organizations around some units. Two children with an equal number of neural connections might vary enormously in intelligence, due to differences in the system of organization of their connections caused either by heredity or experience.

3. *The ease of forming and breaking neural connections.*—Intelligence appears to be highly correlated with plasticity, and plasticity applies equally to the forming or breaking of a connection. It seems to be a general characteristic of stupid individuals that they form connections or learn very slowly, but when a habit of acting or thinking has been established by dint of much effort, they do not readily relinquish or modify it to meet novel or changing conditions. This barnacle-like clinging to old habits should not be mistaken for a good memory. Quick forming and rapid modifications of neural connections are the undoubted attributes of superior intelligence, but quick modification occurs only when experience demonstrates that an old neural connection inadequately adapts the individual to his environment.

4. *Permanence of desirable neural connections.*—It has just been pointed out that intelligence is closely correlated with ease of forming and breaking neural connections. Great permanence of desirable connections is also an index of

high intelligence. No great growth of intelligence could occur if each night of sleep wiped out the learning of the day as each night in Valhalla healed the wounds of the day's battle. There is a popular belief, fallaciously transferred from bank accounts to individuals' memories that "easy comes, easy goes," and hence superior intelligence cannot possess both superior plasticity and superior permanence. Not only "to him that hath shall be given" but to him that hath *has been* given. For the adage that he who learns quickly forgets quickly is not based on facts but upon a sympathetic desire to comfort stupid folks.

This somewhat theoretical analysis is followed by a set of more tangible working principles. Most of these principles have consciously or unconsciously guided those who have constructed our basic intelligence tests.

Suggestions for Intelligence Test Construction.— 1. *An intelligence test should be a learning test which extends backward rather than forward.*—These four—the number of desirable neural connections, the organization of these connections, the ease of forming and breaking connections, and the permanence of connections—are the chief characteristics of an individual which a test must measure if it is to be a good intelligence test. Two methods have been proposed for testing these pre-eminently valuable characteristics. One method is to confront a pupil with a learning situation which varies from very simple to very complex. A measurement of the number of points learned, the maximum complexity of the thing that could be learned, the rate of learning, and the persistence of the things learned, would give a measure of the pupil's four prime characteristics. The inherent difficulties in conducting such learning tests are so great that another testing method is in almost exclusive use. This method takes samplings from the abilities which a pupil has, during his whole life, succeeded in developing. While this is also a learning test method, the learning test extends all the way from the present back to birth rather than from the present to a brief future.

A single type of test will measure all four neural characteristics. Imagine two pupils of like chronological age. The pupil who forms and breaks neural connections more easily and whose connections are more permanent, will have the greatest number of neural connections. Hence, to compare the intelligence of the two pupils, all that is necessary is to make a determination of the relative number of their neural connections or mental abilities and of the efficiency of organization of these abilities or connections.

2. *An intelligence test should measure the largest possible number of traits.*—While it is probably true that every mental trait or set of neural connections boasts no aristocratic exclusiveness, but exemplifies even Nature's predilection for democracy by partially combining with other traits to constitute a coördinating neural hierarchy, nevertheless, every trait retains a portion of its individuality or exclusiveness! For this reason, the larger the number of traits measured, the safer the diagnosis. A test which measured but a few traits might happen to strike just those mental functions in which the pupil, for some accidental reason, was specially strong or specially wanting. The assayer takes many samples from many points in the ore bed.

3. *An intelligence test should measure samplings from the relatively more differentiating traits.*—The ideal way is to measure every trait that contributes to intelligence, attach weights to the various traits according to the amount of their contribution to intelligence and add. The resulting sum would be a perfect measure of intelligence. Even if we knew how to test every trait and just what weights to attach to each, time would compel us to confine our attention to the more differentiating traits.

But how may we know what are the differentiating traits? Let us proceed by the process of elimination. We can eliminate those traits in which man is little or not at all superior to the animals. Certain elemental functions such as keenness of vision, hearing, and smell, speed of simple muscular responses like running and tapping and the excellence of the

neural functioning in connection with breathing, digesting and other organic functions, all these are of great importance. There are more failures from indigestion than this world dreams of. After a certain minimum these traits have little differentiating value. In them the brute and the stupid human are about the same as the genius. The intelligence tester will do well to steer clear of simple sensori-motor tests and seek out those traits which chiefly distinguish man from the brute, and genius from stupidity. Simple observation of these distinctions will point the way toward differentiating traits.

Both observations and correlations with semi-satisfactory estimates of intelligence have indicated that intelligence tests should measure such traits as the ability to analyze a complicated situation, to attend to many elements at one time, to easily and effectively shift from one mental set to another, to deal with abstract symbols and relationships and the like. I [4] found that complex, difficulty tests show a closer correlation with intelligence than simple speed tests such as cancelling figures or letters, adding, copying and the like. The preceding psychological analysis would lead to just such a conclusion because it is the complex difficulty which indicates the efficiency of organization of a pupil's neural connections.

It is not sufficient to select and test the differentiating traits. At best the traits tested will vary in differentiating power, or said differently, will vary in the amount of their contribution to intelligence. Thorndike [5] has pointed out that refined procedure requires each trait to be weighted according to its contribution to intelligence.

Theoretically the weight attached will often differ not only from trait to trait, but will differ for different amounts of one trait. Just as one trait may contribute more to intelligence than another, so the same trait may not contribute anything

[4] Wm. A. McCall, "Correlations of Some Psychological and Educational Measurements with Special Attention to the Measurement of Mental Ability"; *Teachers College,* Columbia University Contributions to Education, No. 79.

[5] For an excellent discussion of how to weight traits see E. L. Thorndike, "Fundamental Theorems in Judging Men"; *The Journal of Applied Psychology,* March, 1918.

until it is present in considerable amount, after which each increase may contribute a proportional amount of intelligence until a certain point is reached beyond which further increases in the trait are attended by a decreased contribution. When a man is seeking a wife, intelligence itself becomes a trait contributing to wife-likelihood for any woman he meets. It is said that any increase in a woman's intelligence contributes nothing toward a man's desire to marry her until a certain minimum of intelligence is reached. Any increase in intelligence from this minimum to a certain maximum, usually defined as on a par with the man's intelligence, makes her more desirable to him. But woe to the woman whose intelligence exceeds his maximum unless she is clever at camouflage! Beyond that maximum an inverse relationship obtains, or so it is said. Most of the scientific work that has been done with mental traits has, however, assumed a constant and not a changing relationship such as that of the foregoing illustration. Thorndike states in the above mentioned article:

"This assumption of proportionality or rectilinearity in the relation line is behind most of the scientific work that has been done with educational and vocational selections. The technique of partial correlation coefficients and the regression equation, for example, assumes approximate rectilinearity of the relation lines."

In attaching weights to traits the examiner should be specially careful lest he weight the same trait more than once. Suppose that we are assigned the task of selecting promising salesmen for life insurance. Suppose that among other traits we have tested language ability and mathematical ability, and that we wish to assign appropriate weights to each ability. If, due to complicated test instructions or roundabout descriptions of problems, the chief difficulty in the so-called mathematical tests were really language difficulty and not mathematical difficulty, assigning weights to the two tests would thus really mean the assignment of a

double and hence an undue weight to language ability. This error of weighting the same trait more than once is very common. The error can be avoided only by a knowledge of each test's correlations. Even though two tests have different names, if the correlation between them is very close, we can be pretty sure that the difference is in name only. To quote from the same article by Thorndike:

"Do not weight the same contributing element twice because it appears in two or more traits. Or, more adequately: Attach weights to elements according to the amount of their contributions, irrespective of the number of symptoms in which they appear. A full determination of independent elements and their contributions, whether experimental or by the method of partial correlation coefficients, is very difficult, and has never been made in any case to my knowledge. A complete determination is indeed not necessary for fairly efficient prophecy. Where the best possible weighting would prophesy fitness with say, a resemblance to demonstrated fitness of .95, a prophecy to the extent of .90 will commonly be reached by such a rough weighting as a competent thinker can devise in an hour or two upon inspection of the relevant correlations. So for practical purposes, we may translate the theorem into: Attach weights to traits only after a knowledge of their correlations."

The examiner's object should be to test traits which are each closely correlated with intelligence, but not closely correlated among themselves. Only thus is it possible to avoid a lop-sided view of intelligence.

Weights should be attached to a trait not only in the light of its contribution to intelligence, and in the light of variation in contribution with variation in the trait's amount and in the number of symptoms in which the trait appears, but also in the light of the trait's influence as affected by the amount of presence or absence of other traits upon which the trait in question depends. To illustrate, an increase in intelligence does not cause an increase in demand for one's services as a bank teller if one is dishonest. But let there

be an increase in the trait honesty, and then increases in intelligence otherwise valueless become valuable. With the explanation that Thorndike is talking about a man's ability for a job rather than a pupil's intelligence we cannot do better than let him state this point and summarize the whole discussion. (It should be said that these principles apply to the measurement of any ability which is broad enough to be dependent upon subordinate traits.)

"The dependence of the amount of influence of a trait upon the amount of some other trait possessed by the man will often be more complex than that described in the illustration. For example, the value of technical knowledge and skill for a certain job might vary as the square root of the man's intellect or as the square of his intellect. There may also be many sorts of irregular relations in these dependencies of one trait's influence upon the amount of another trait that is possessed.

"I am not aware that any scientific investigation has measured any of these dependencies. There is, however, good reason to believe that they exist and are important. A man's possession of what we call energy, for example, seems to be a multiplier for his intellect or skill. A man's loyalty or devotion in any particular job seems to be a multiplier of his other equipment for it. What we call roughly interest in success, or determination to succeed, or ambition, seems to be a multiplier for energy. And in theory, at least, we must admit it is a fundamental theorem that prophecy of the degree of influence of any amount of a trait requires consideration of the amount of every other trait in the man in question.

"There are other principles concerning the use of facts in the selection of men which must be followed to make the best possible prophecy, but limitations of time prevent their discussion now. In the time that remains I may simply summarize our findings roughly and compare the scientific with the impressionistic or intuitional use of facts about a man.

"We have seen that the status of a man in the traits relevant to fitness for a job may be expressed in an equation:

John Doe = 7a + 9b + 4c + 21d. . . . A prophecy of his fitness relative to other men is obtained by attaching weights to a, b, c, d, etc., in view of (1) their relations to fitness, (2) their partial constitution by common elements, and (3) any dependencies whereby one gains or loses in influence according to the amounts of the others which are present. The setting up of an equation of prophecy from an equation of status will usually be very complex, but a rough approximation, if sound in principle, will often give excellent results. In so far as the lines of relation, interrelation, and dependency are rectilinear, the technique is greatly simplified; and a rough approximation to this is probably often the case.

"Even when approximations are used and when, by good fortune, many of the relations concerned are rectilinear, sound procedure will still be more elaborate and difficult probably than has ever yet obtained in practice. Practice is thus justified to some extent in judging fitness experimentally by tests with dummy operations like those of the job in question, as well as analytically by a weighted combination of contributing elements.

"All this study of relation lines, intercorrelations, facilitations and inhibitions and resulting weights by multiplying and adding, represents a scientific execution of just what the competent impressionistic or intuitional judge of men tries to do. The strength of such intuitional judgments in comparison with the formal systems of credits and penalties of the past has been, not that the intuition was less quantitative, but that it was more so! The formal systems of the past have used symptoms merely additively and often with only an 'all or none' credit. They have not allowed for the undue duplication of credits by intercorrelations, and have not sensed the importance of the multiplying effect of certain traits upon others. The competent impressionistic judge of men does respond to these interrelations of the facts and sums up in his estimate a consideration of each in the light of the others. If there are ten traits involved, say ten entries on an application blank, he may be said to determine his prophecy by at least 10 + 9 + 8 + 7 + 6 + 5 + 4 + 3 + 2 + 1 quantities, since he responds to each trait in re-

lation to all the others. There is a prevalent myth that the expert judge of men succeeds by some mystery of divination. Of course, this is nonsense. He succeeds because he makes smaller errors in the facts or in the way he weights them. Sufficient insight and investigation should enable us to secure all the advantages of the impressionistic judgment (except its speed and convenience) without any of its defects."

4. *An intelligence test should measure only those traits which every pupil has an equal opportunity to develop.*— This means that the test material and methods of the test should be drawn from the social medium common to all children. Theoretically there should be *equal* opportunity to learn the test material, but practically about all that can be provided for is *ample* opportunity. Those traits should be measured which are least influenced by such differential agencies as school *vs.* non-school training, city *vs.* rural life, masculinity *vs.* femininity, luxury *vs.* poverty, etc. A country boy might easily show up unfavorably in comparison with his city cousin in reacting to questions about elevators, skyscrapers, subways and roller-skates, while the situation might be reversed if the questions dealt with hay-mows, disc-harrows, silos, dibbles, copperheads, and yellow jackets. Many of our pedagogical tests are increasing in value as intelligence tests because schooling is coming more and more to be a common environment for all children.

5. *An intelligence test should measure those traits growth in which is most universally motivated.*—There are numerous situations in which all children move and have their being and which, according to a liberal definition of environment, are a common environment to all children. Yet many of these situations make little or no appeal to a child unless he has an erratic nature or is on some way extraneously motivated.

If a test incorporated such situations it would unfairly favor children with peculiar interests or peculiar training. The best guarantee of universal motivation is to select for

test elements, activities in which the largest number of children are instinctively interested. The accumulating records of observations on what activities are self-chosen by the children in the kindergarten and primary grades offer a rich mine of suggestions to makers of intelligence tests.

6. *An intelligence test should show a higher per cent of correct responses with each increase in chronological age.*—This principle holds up to the age when intelligence matures, which is supposed to be not far from 18 years of age. The fundamental assumptions underlying the intelligence test as customarily used are that the total amount of knowledge, skill and power acquired by an individual (a) is a measure of his present intelligence, (b) is proportional to his inherited intelligence, and (c) is prophetic of his future intelligence. These three points mean that if one infant has a native endowment twice that of another child, he will develop proportionately faster until intelligence matures and hence will at every stage of his life be proportionately superior to the originally inferior individual.

7. *An intelligence test should measure the ability to transfer training.*—One of the great advantages of possessing numerous neural connections which are effectively organized is that they guarantee wide-scale transfer. The genius makes everything grist which comes to his mill. He can transfer both Latin and Algebra to just about anything. The stupid individual, on the contrary, lacks this nimbleness of wit. He can be trained but can be educated only with difficulty. He would make a fairly good showing if the test contained material upon which he had had direct training. When the test presents tasks for which he has had no specific training his existing neural connections are unable to deal with the new situations. This difference between individuals in their power to deal with situations for which they have had no specific training is so significant and marked that intelligence might well be defined as the power to transfer training. An intelligence test which does not measure this ability is certainly imperfect.

Current Methods of Measuring Intelligence.— If the reader will think of the following quotation from Terman [6] as applying to pupils instead of to soldiers and adults, he will have an admirable statement of current methods of measuring the general intelligence of school children.

"The Method of Trying Out.—One method would be to try out each soldier in tasks of various degrees of difficulty. The method is fairly sure. It could probably be depended upon to give us an efficient army within a few years.

"So with the method of natural sifting. If the war should last long enough the best men would pretty certainly in time demonstrate their ability and rise to positions of responsibility, while those of poorest mentality would gravitate to the humbler tasks. But the gravitational method of sifting meets many resistances, and like the method of trying out, is necessarily very slow.

"Pseudo-scientific Methods of Rating Mentality.—A century ago a French physiologist, by the name of Gall, founded what he thought was a new science, which was named phrenology. According to phrenology, definite and constant relations were believed to exist between certain mental traits and the contour of the head. It was believed, for example, that one's endowment in such traits as intelligence, combativeness, sympathy, tenderness, honesty, religious fervor, and courage, could be judged by the prominence of various parts of the skull.

"It is unnecessary to question the sincerity of Gall and his over-enthusiastic followers. They were probably not guilty of conscious deception, but merely blinded by an attractive theory. At any rate, the 'science' of phrenology has been hopelessly exploded. It has been well demonstrated:

"1. That traits like those mentioned above do not have separate and well-defined seats in the brain, and

"2. That skull contour is not a reliable index of the brain development beneath.

"In the underworld of pseudo-science, however, phren-

[6] Lewis M. Terman, "Tests of General Intelligence," *Psychological Bulletin*, May, 1918.

ology and kindred fakes still survive. Hundreds of men and women still earn their living by 'feeling bumps on the head,' reading character from the lines of the hand, etc. But modern warfare has no time for pseudo-science. A general would no more think of selecting his officers by phrenological methods than of substituting incantations for gunpowder.

"The Method of Off-hand Judgment.—But if in the rating of men pseudo-science is misleading, perhaps science is still unnecessary. It may be argued that mental traits can be rated accurately enough for all practical purposes on the basis of ordinary observation of one's behavior, speech and appearance. We are constantly judging people by this off-hand method, because we are compelled to do so. Consequently we all acquire a certain facility in handling the method. For ordinary purposes it is infinitely better than nothing. A skillful observer can estimate roughly the height of an aëroplane; but if we would know its real height we must use the methods of science and perform a mathematical computation.

"The trouble with the observational method is its lack of a universal standard of judgment. One observer may use a high, another a low standard of comparison. A four-story building in the midst of New York's 'sky-scrapers' looks very low; placed in the midst of a wide expanse of one-story structures it would look very tall. The captain of a very superior company may rate his least intelligent man as 'very dull'; the same man in a very inferior company would likely be rated as 'average' or better.

"Moreover, we are easily misled by appearances. The writer knows a young man who looks so foolish that he is often mistaken by casual acquaintances for a mental defective. In reality he is one of the half dozen brightest students in a large university. Another man who in reality has the mentality of a ten-year-old child, is so intelligent looking that he was able to secure employment as a city policeman.

"Language is a great deceiver. The fluent talker is likely to be over-rated, the person of stumbling or monosyllabic speech to be underrated. Similar errors are made in judging the intelligence of the sprightly and the stolid, the

aggressive and the timid, etc. Our tendency is also to overestimate the intellectual quality of our friends and to underestimate that of persons we do not like.

"If the method of off-hand judgment were reliable, different judges would agree in their ratings of the same individual. When the judges disagree it is evident that not all can be correct. When intelligence is rated in this way wide differences of opinion invariably appear. Twenty-five members of a university class who had worked together intimately for a year were asked to rank the individuals of the class from 1 to 25 in order of intelligence. The result was surprising. Almost every member of the class was rated among the brightest by someone, and almost every member of the class among the dullest by someone. Doubtless the judges were misled by all kinds of irrelevant matters, such as personal appearance, fluency of speech, positiveness of manner, personal likes and dislikes, etc. Think how much error there would be if a company commander were rating the intelligence of 250 men newly assigned to his command.

"The method of personal estimate is much better than the method of external signs (phrenology), but to be reliable it must be supplemented by a method which is *objective,* that is, a method which is not influenced by the personal bias of the judge or by such irrelevant factors as appearance, speech, or bearing of the one to be rated. Such is the method of intelligence tests.

"Intelligence Tests a Method of Assaying Mentality.—A man wishes to find out the value of a gold bearing vein of quartz. How shall he set about it? One way would be to uncover all the ore and extract every ounce of gold contained in it. It is hardly necessary to point out that this would be a slow and risky procedure, one that might easily cost a fortune and bring small returns. But granting that the extent of the quartz vein was known and that the cost of bringing it to the surface could be calculated, would this be sufficient to tell us the value of the mine? The answer is obvious; something depends on whether the quartz contains many dollars' worth of gold or only a few pennies' worth, per ton of ore.

"However, the next step is easy. It is only necessary to take a few random samples of the ore to an assayer, who makes a simple test and returns the verdict of so many ounces of gold per ton of rock. The verdict of the assayer may justify the expenditure of a million dollars or it may tell us the mine is not worth a penny. At any rate the question of value is answered.

"Suppose the question before us is not the value of a gold bearing vein of quartz, but the intellectual quality of a human mind. If we are to rate the quality of a man's intelligence will it be necessary to make this intellect perform every act of which it is capable in order that these may be added together for a total intelligence rating? This would be one method of answering the question, but a rather tedious one, considering the innumerable acts which a human mind is able to perform. Perhaps this is not necessary. Conceivably it might be possible to sink shafts, as it were, at certain critical points, and by examining a few samples of the mind's intellectual product to estimate its intrinsic quality by a method analagous to that of the assayer.

"Such is the method employed in all systems of testing intelligence. The mind is given a number of 'stunts' to perform, each of which requires the exercise of intelligence. By the quality of these the quality of the entire mind is judged. The tests tell us whether the mind in question is one of rich content and rare intellectual power, or whether it is mediocre or perhaps even defective.

"Collecting Samples for Assaying.—In ascertaining the value of the gold deposit would it be safe to take all the assayer's samples from a single part of the quartz vein? Common sense would of course suggest the precaution of taking samples from many places and of estimating the gold content in terms of average richness. Similarly in testing intelligence the subject is not asked to perform one intellectual 'stunt,' but many. He may be given tests of memory, of language comprehension, of vocabulary, of orientation in time and space, of ability to follow directions, of knowledge about familiar things, of judgment, of ability to find likeness and differences between common objects, of arith-

metical reasoning, of resourcefulness and ingenuity in practical situations, of ability to detect the nonsense in absurd statements, of speed and richness of mental associations, of power to combine related ideas into a logical whole, of ability to generalize from particulars, etc. The average of a large number of performances thus gives a kind of composite picture of the subject's *general intelligence.*"

CHAPTER VIII

ORGANIZATION OF TEST MATERIAL AND PREPARATION OF INSTRUCTIONS

I. Factors Which Should Influence Arrangement

Test Forms.—An important requirement in test construction is that the scoring be as economical, accurate and objective as possible. There are two ways of meeting these requirements. The first way is the proper construction of the test, the second way is the proper construction of scoring devices.

In test construction the prime requisite from the point of view of scoring is that those pupil reactions to the test which are to be scored be as simple, abbreviated and controlled as possible, and that the reactions have a definite spatial location. With the exercise of some ingenuity a pupil's most complicated mental processes can be measured even when he reacts to each test element with no more than a word, a letter, a check, a number or the like. The excellence of the pupil's solution of a long reasoning problem in arithmetic can be condensed into a few figures—the answer. If the pupil's reactions are simple and abbreviated they can be scored very rapidly and accurately, and with very little disagreement among the scorers.

Again, a test must also so control these reactions that only one kind of simple reaction will be correct. If any one of ten different words, or letters or numbers is correct, scoring will be greatly slowed up and judgment must be more and more exercised and the net result is uneconomical, inaccurate, and subjective scoring. If only one reaction is correct for a given test element it is possible to make out a set of correct answers. These correct answers may be placed

beside a pupil's answers, and then scoring becomes merely a matter of making simple, unthinking, visual comparisons.

Finally, the test must be so constructed as to give a definite spatial location to a pupil's answers. In any case this is a decided convenience; it is particularly so when a pupil's reactions all consist of a check mark or an underlining, where correctness depends not so much upon what is done as where it is done. Spatial location is secured by the provision of a square, circle, or other special place where the pupil is to make his mark. Consider, for example, how long it would require to announce the results of a presidential election if ballots did not spatially locate the voter's vote.

The problem of constructing a test so that scoring will be efficient is shown by the following evolution of an extract from a test for military aviators which the writer aided Thorndike in constructing. (Instructions are omitted.) Note first that the nature of the test question permits a long, qualified, unscorable answer. Note second that there is no prescribed place where the answer must be written. This test element is a perfect illustration of what not to do.

1. Compare the lines as they were before with what they are now.

The test element is restated in better form below, though it is still inexcusable. Note that the nature of the test element encourages a briefer answer, and tends to control the type of answer.

1. Are the lines shorter than they were before, longer than they were before, or the same as they were before?

The test element is restated again in a still better form. the aviators were instructed to write the appropriate number in the parenthesis as I have done in the illustration. Note that the answer is simple, abbreviated, controlled, and located somewhat apart from the statement of the question.

1. Are the lines (1) shorter than they were before, (2) longer than they were before, or (3) the same as they were before? .... (2)

The above is the first form in which the question was actually stated. Note that a column of correct answers, properly spaced, could be placed beside a column of an aviator's answers in such a way that all errors could be detected with great accuracy and rapidity.

But there is a decided defect in even the above form. Suppose the lines or trenches really are (2) i. e. longer than they were before. For the aviator to report to the Intelligence Officer that the lines are shorter than they were before is to make a more serious mistake than if he were to report that they are the same as they were before. The former should be penalized, say, two points and the latter only one point. Consider how the following re-arrangement facilitates the assignment of the proper amount of penalty.

1. Are the lines (1) shorter than they were before, (2) the same as they were before, or (3) longer than they were before? .... (3)

Since in this case the answer should be the lines really are longer than they were before, the correct answer is 3. If 2 is found in the parenthesis, it should be penalized 1 point. The difference between 3 and 2 is 1 point. If 1 is found in the parenthesis, it should be penalized 2 points. The difference between 3 and 1 is 2 points. Thus the test element has been so constructed that the difference between the correct number and the number appearing in the parenthesis gives instantly the proper amount of penalty. Without such simplification of scoring the extensive use of mental tests in the military service during the war would not have been possible, nor would there be great promise for their future use in education.

Below are extracts from a variety of tests, which illustrate how not only tests but ordinary examinations can be so con-

structed as enormously to reduce the inaccuracy, subjectivity, and time of scoring.

EXTRACT FROM RUGER'S PROVERBS TEST

DIRECTIONS: *In column No. 1 write opposite each English proverb the number of the African proverb which most nearly means the same thing as the English proverb (see below for African proverbs). (Do not write any number twice—omit no number—write only one number opposite each letter.)*

Column 1	2	English Proverbs
	a.	First catch your hare.
	b.	Curses come home to roost.
	c.	Milk for babes.

AFRICAN PROVERBS

1. Ashes fly in the face of him who throws them.
2. I nearly killed the bird. No one can eat nearly in a stew.
3. If the stomach is not strong, do not eat cockroaches.

EXTRACT FROM THORNDIKE'S MENTAL ALERTNESS TEST

Make a cross in the square before the best answer to each question.

1. Why are prunes a good food? Because they
   - [ ] grow in California
   - [ ] are wholesome and economical
   - [ ] are served in boarding houses
   - [ ] make an attractive dish

4. When you feel that affairs in your town are badly managed, should you?
   - [ ] do nothing at all
   - [ ] growl to your friends
   - [ ] get out and work to have things changed
   - [ ] go to church

EXTRACT FROM PRESSEY'S MENTAL SURVEY TEST[1]

X. Analogies.

Examples:
girl—woman: boy—man
sun—day: moon............................
good—bad: big............................

1. woman—girl: man..............
2. kitten—cat: puppy............
3. sky—blue: grass.............

11. hill—valley: high.............
12. arm—elbow: leg..............
13. truth—falsehood: straight line.................

[1] Issued by S. L. and L. W. Pressey, University of Indiana, Bloomington, Ind.

EXTRACT FROM GREENE'S ORGANIZATION TEST[2]

Write numbers in these spaces

1. (1) a dog, (2) a boy, (3) had..............	
2. (1) of the cold, (2) afraid, (3) they were....	
3. (1) I am, (2) see, (3) how tall.............	

EXTRACT[3] FROM OTIS' GROUP INTELLIGENCE SCALE

MEMORY

DIRECTIONS: *Read each question and if the right answer, according to the story, is* YES *draw a line under the word* YES. *If the right answer is* NO, *draw a line under the word* NO. *But if you do not know the right answer, because the story didn't say, draw a line under the words* DIDN'T SAY.

Sample:
- Was the story about a king? — yes no didn't say
- Was the king's daughter sixteen years old? — yes no didn't say
- Was she ugly? — yes no didn't say

Begin here:

1. Was the king fond of hearing stories? (yes no didn't say) 1.
2. Did the king offer his daughter to any one who could tell him a story that would last forever? (yes no didn't say) 2.
3. Did he offer all his kingdom also? (yes no didn't say) 3.
4. Did he say, "but if he fails he shall be cast into prison"? (yes no didn't say) 4.

Mechanical Scoring Devices.—Since scoring is greatly facilitated by mechanical scoring devices and since the possibility of employing such devices is dependent upon the form of arrangement of the test material, a brief discussion of these devices is pertinent at this point.

There are many forms of these mechanical devices depending upon the form of the test which they are designed to score. When all the pupils' answers are written at a definite place on the right or left edge of the test sheet a convenient device is a test sheet which has been correctly filled out by the scorer. The key sheet can be so superimposed on the

[2] Issued by S. A. Courtis, 82 Eliot Street, Detroit, Mich.

[3] Copyrighted 1919 by World Book Company, Yonkers-on-Hudson, New York. Used by permission of publishers.

pupil's sheet that nothing but the pupil's column of answers shows immediately beside the correct answers. Such a scoring device can be used even when the pupils' answers are written between the lines. A test sheet is correctly filled and then all but the answers are cut away. The pupils' answers show through the resulting spaces, or the scoring sheet may be rolled and unrolled up and down the page. It is well to use a test sheet for the scoring device, because it quickly and automatically provides for proper spacing. If a more durable and less flexible form is required, the key answers may be placed on a card-board or even more durable material.

Again, there are tests of such a nature that what the pupil does is relatively insignificant but where he does it is all-important. Such are tests where the pupil is instructed to underscore the appropriate word, or check the appropriate reason, or cancel the appropriate letter, etc. The scoring device already described may be used to advantage in this situation, but some form of transparent sheet frequently works better. Celluloid or any kind of transparent material may be placed over a correctly filled test, and a dot can be made on the celluloid sheet just over the place which is correct. The transparent sheet may then be used for scoring the pupils' answers. Otis makes an extensive use of just such a device for scoring his group intelligence test.

Finally, if the test is so constructed that scoring will be facilitated by making all of the pupil's test sheet invisible except the spot where the correct answers should be, small apertures may be cut through a blank sheet at such places that only the correct-answer spots will be visible. The same result may be secured by placing a sheet of celluloid over a test sheet and by so painting the celluloid with black paint that nothing but the desired spots will be visible. These perforated scoring devices may also be used to facilitate the counting separately of items mixed in one test. The Woody-McCall Fundamentals of Arithmetic Test,[4] Form I and

[4] Bureau of Publication, Teachers College, N. Y. C.

Form II, has addition, subtraction, multiplication, and division examples so mingled on one test sheet that the pupil is frequently forced to shift his processes, and often to decide by the nature of the sign just what sort of an example it is. In the instructions which accompany this test, I suggest that the computation of a separate score for each fundamental, if desired, may be facilitated by perforating four fresh test sheets. The first sheet should be so perforated that when placed over the pupil's test only addition examples are visible. The second sheet should make visible only subtraction examples; and multiplication and division should be treated similarly.

Group *vs.* Individual Testing.—The nature of certain tests and the illiteracy of young pupils has required individual testing, i. e., the testing of one pupil at a time. The nature of other tests and the literacy of older pupils permits group testing, i. e., the testing of many pupils.

A heated controversy has been going on concerning the advantages and disadvantages of each method of testing, and this controversy continues in spite of the fact that skillful test constructors have now adapted almost all varieties of tests to permit group testing of illiterates.

Even when group testing is feasible, it is claimed that a more accurate diagnosis can be made when each pupil is tested individually. This claim is based upon the assumption, first, that the appearances and incidental reactions of a pupil are valuable indices of his special defects or special strengths and that these indices are observed better during an individual examination. The second assumption is that the examiner can better select for each pupil those tests which will reveal significant symptoms, for it often happens that some reaction on the part of the pupil will give the examiner a "lead" which it is highly desirable to follow up. Such rapid adaptations are manifestly impossible in group testing. Finally, some examiners hold that testing conditions can be more carefully standardized by individual testing. Early psychological investigators considered them-

selves unusually virtuous when they took time to administer all tests individually "with special care," as they said.

Group measurement is enormously economical in time. To administer a thirty-minute individual test to a group of 500 pupils would, when all wastage is counted, take about 300 hours of the examiner's time, whereas, under certain circumstances, a thirty-minute group test could be administered to all pupils in about forty minutes. Even though the 500 pupils were tested in groups of only fifty, a great saving of time would be effected. It is this great expense in time that has delayed educational measurement in the kindergarten and primary grades. The economy of group testing is further illustrated by the psychological examination of soldiers during the war. Several tests were given to many hundreds of thousands of soldiers. Each test could have been administered individually to each recruit. To have done so with the staff available would have required all the years of the war, when speed was imperative. Substantially the same situation confronts those who are introducing measurement into education. It is useless to attempt the measurement of millions of pupils with individual tests. To a very large extent educational measurement must be group measurement.

Group testing may, under certain conditions, be fairer to the pupils tested. In experimentation it is often important to know the amount of change made by each pupil in a class during two weeks. It might take a single examiner a week to test every child by the individual method. The last pupils tested would thus have an extra week's advantage if learning were being measured, or a week's disadvantage if forgetting were being measured. Again, a test is often of such a nature that one pupil can partially prepare another. The first pupils tested can then spread information through the entire class or school. Finally, for some tests, it is especially difficult to standardize the personal equation of the examiner. Such a variable operates to the advantage of some pupils and to the disadvantage of others. Group test-

ing makes the personal equation more nearly constant for all pupils within the group.

What then is the conclusion of the whole matter? Individual testing and group testing each secure special values. The method adopted in the psychological examination of soldiers will probably come into common use in all educational measurement whether done for purely pedagogical or clinical purposes. The initial tests given the soldiers were group tests. These revealed the illiterates and those who were in some way abnormal. The illiterate and abnormal groups were then intensively measured with individual tests. The diagnoses afforded by the group tests were accepted for the vast majority of the recruits. In time school psychologists will not wait until abnormal cases are sent to them for diagnosis. They will sweep through the schools with a net of group tests and catch their own cases for intensive study. Even for the special cases, what with the development of group tests for illiterates, it is worth considering whether the greater number of group tests which may be given within an equal time-interval may not give a better diagnosis than the fewer individual tests. A good practical rule is to *first give group tests, accept their diagnosis for most of the pupils and give further group or individual tests to the few pupils, who, according to the group tests, need special study.*

II. Guiding Principles in the Preparation of Instructions

1. ***Instructions Should Be as Brief as Is Consistent with an Adequate Understanding of What Is to Be Done.***—Besides consuming time, inordinately long instructions tend to produce confusion in the minds of the pupils. Even adults find difficulty in following through complicated instructions. It has been demonstrated frequently that even among so intelligent a group as school teachers there are always a few who cannot follow very simple directions. Long instructions so tax the memories of pupils that

absolute essentials are frequently forgotten. To forget a single one of these essentials may markedly alter the child's score. Brevity is frequently sacrificed to pure irrelevancies. It is well to remember that the primary function of instructions is to give a pupil adequate, but not necessarily complete, information about the test. Their primary function is not to give the pupil a general education. To quote a remark by Patterson, "Test! Don't teach!"

Again, the longer we make the instructions, the more we add to the confusion of inexperienced examiners. The novice is never quite sure of himself unless the instructions are sufficiently brief that his memory span can embrace not only every step of the process, but also the proper sequence of the steps. The untrained examiner cannot give his sole attention to instructions. He must maintain order among a roomful of naturally disorderly creatures, keep track of his watch, handle the test sheets, see that preceding instructions are being followed and the like. It is a real kindness to both examiner and pupils to make instructions no longer than is necessary.

But inadequate instructions are as bad as or worse than instructions which are too long. Inadequate instructions may wholly defeat the purpose of the test, or precipitate an avalanche of questions from the pupils. Instructions cannot be cut out of whole cloth. It requires both forethought and experimentation to produce instructions which will cause the pupils to do just what is wanted of them, and which will anticipate questions by the pupils.

The omission of some points would be more disastrous than others. What the essential key points are depends, of course, upon the test. In the Thorndike Vocabulary Scale, for example, it is especially important that pupils be warned not to skip any words by accident. This is because the statistical method of computing scores for this test treats accidental omissions as though they were errors, and weights them very heavily. Below are a few quotations

from existing test instructions which are key points. "As soon as you complete the first sheet, hold up your hand, and I'll give you a second one." "Read as rapidly as you can to still understand what it says." "Don't read anything over again." "You will have just one minute." "This is an addition test." "Check each sum before passing to the next example." "When I call 'stop,' draw a circle around the last word read." "You will be asked to reproduce from memory what you have read." "Your score will be the number of examples you get right." "You will be marked on both speed and quality." "Write your name and grade." Some key points are so obvious that they will be recognized by anyone. Some are so subtle that only the intuitive or trained examiner can detect them. In sum, instructions should be as brief as possible, as adequate as is essential, and always consistent with the subsequent uses of results.

2. ***Instructions Should Employ a Demonstration and Preliminary Test.***—An ounce of demonstration is worth a pound of words! It takes more words to describe effectively what is to be done than it takes moves to show what is to be done. Anyone can try for himself an experiment to discover whether it is easier to show than to tell. Probably due to primordial practice, children, not to mention adults, can imitate better than they can comprehend and follow linguistic directions. To accompany description with a demonstration not only caters to pupils who may get impressions easier through the eye or through the ear, but, what is more important, it gives to all an impression through both eye and ear. Demonstration has the still further advantage of securing better attention, especially from the young children.

The demonstration may take any of several forms. In one test the examiner writes a sample test element on the blackboard and works it out for the pupils just as they are to work out similar tasks contained in the test. But in most tests which employ the demonstration method, sample test

elements correctly completed are printed on the test sheet. Here is an example of instructions for a test accompanied by such a completed sample.

"This is a test of common sense. Below are sixteen questions. Three answers are given to each question. You are to look at the answers carefully; then make a cross in the square before the best answer to each question, as in the sample:

SAMPLE

Why do we use stoves? Because

☐ they look well

☒ they keep us warm

☐ they are black

"Here the second answer is the best one and is marked with a cross. Begin with No. 1 and keep on until time is called."

Thorndike has devised a novel test. This test attains the maximum of showing and the minimum of linguistic directions. So much is this the case that it may well be called a *pantomime* test. The whole test can be given without the reading or the speaking of a word by anyone. The test was devised, in fact, to measure the intelligence of army recruits who were illiterate Americans and immigrants who did not even understand spoken English. The recruits were given a test sheet containing diagrams, pictures, etc. The examiner placed before the recruits an enlarged form of the test which was similar to, but not identical with, the test in the hands of the recruits. The examiner did the enlarged test with a heavy crayon. The examiner's movements showed the recruits what they were to do with their own test sheet. This is a most ingenious test, but, when there is a common medium of communication, the best method of giving instructions is not by demonstration alone, nor by linguistic description alone, but by a happy combination of both.

When instructions are at all complex, they should, as a rule, be accompanied by a preliminary test. Even though every possible precaution be taken to make all pupils under-

stand just what they are to do, one can never be quite sure that all do understand unless a preliminary test is given. A preliminary test has the additional advantage that pupils can make most of their test adjustments before beginning the test proper. Due to differences in nervousness, intelligence, etc., some pupils adjust quickly and some slowly. If there is no preliminary test, and if the time for the test is relatively brief, the rate of adjustment may materially influence the score, even when we are usually not primarily concerned with the measurement of this factor. The preliminary test should typify the nature of the test elements proper.

This preliminary test may be presented in various ways. Sometimes the examiner writes one or more typical test elements on the blackboard and the children do them more or less in concert. Obviously this method does not give the examiner a sure guarantee that each pupil understands what is expected of him.

A second method is to give each pupil an easy miniature test. The examiner can then go about the room and observe whether each pupil shows an understanding of instructions. The examiner can help any pupil do the first element or two if he does not understand. If this does not suffice, the pupil can be assumed to be incapable of doing the test at all. Such special preliminary tests have not come into general use because of the expense involved in printing extra test sheets and the time required for their distribution and collection.

A third method is to print the preliminary test on the back of the regular test sheet along with the instructions or to reserve the front page of a booklet for instructions and preliminary test. This method is most satisfactory of all. Its use is not universal because of the greater expense involved in printing on both sides of a test sheet or making a booklet.

A fourth method is a little less satisfactory and, as a compensation, less expensive. The instructions, demonstrations, and preliminary practice test can be printed on the same side

of the sheet as the regular test, but clearly separated from the regular test. Pupils can be instructed to do the practice test, but not to begin the regular test until their work on the preliminary test has been inspected and they have received the signal to start the test proper. It is difficult to prevent a premature *mental* start. If the test is a rate test such a premature start may be a serious factor.

A fifth method has been used. When the time element is not important, the elements of the preliminary test may be, so far as the pupil is informed, the first few elements of the regular test. After the test has been started the examiner can go about the room and give any needed help on the preliminary elements. In this case the preliminary elements will not be counted in determining the pupils' scores.

Sometimes practically all the advantages of all the methods can be secured by folding back the preliminary portion of the test in such a way as to conceal the regular test while the preliminary test is visible. This permits printing the test by a single impression, and thus reduces expense. If expense is not, however, a consideration, the folder or booklet test, with the entire front page exclusively reserved for name, grade, age, instructions and preliminary test, is preferable.

Finally, it remains to be pointed out that there are cases where the preliminary test cannot be used at all or its use must be carefully guarded. No one will, of course, make the mistake of using as a preliminary test element, an element which is identical with one in the regular test. To do this would be to teach the test. Again, there are instances where the purpose of the test is to discover whether the pupil possesses the slightest trace of an ability. A preliminary test with the examiner's assistance, or even a demonstration might give sufficient instruction to defeat the purpose of the test, especially if the ability is one which can be quickly taught. This would be true for many elements of the Binet-Simon Test. The common sense of the examiner can be trusted to decide when a demonstration or preliminary test

of a given kind is pernicious. All that is necessary is to keep in mind the fact that demonstrations and preliminary tests are not only clarifying but educating agencies.

3. ***Instructions Should Be Adapted to and Uniform for All Who Are to Be Tested.***—How much adaptation is essential? In the testing of school abilities, the instructions for the test should be so simple that all may understand them. The instructions should be such that no child will fail to make a score just because he failed to comprehend the instructions. Fully to realize this aim, the instructions should be no more difficult than the first or easiest test element. If instructions are more difficult than the first test element, some pupil may make a test score of zero when he might at least have made a small score. In such a case the pupil has not been tested by the test but by the instructions.

How much uniformity is essential? Instructions contain mechanical and non-mechanical features. The mechanical phase has to do with getting the pupil's name, sex, age, grade, etc. Uniformity is not necessary because the important thing is to get this data of identification, even though it is necessary for the examiner to so vary the procedure as to write the pupil's name for him. The mechanical features do not assist the pupil with the test proper.

The non-mechanical features do determine to a certain extent, and frequently to a large extent, the score a pupil will make. It is far more convenient if these instructions are uniform from grade to grade. To cite one illustration, tests are frequently used in rural schools where several grades and many ages are grouped in one room. An examiner can test all these pupils at once if the instructions are uniform. Hence it is best for instructions to be both adapted to and uniform for all the pupils in all the grades.

The intelligence examiner will grumble because I have not been even more enthusiastic for absolute uniformity. The intelligence examiner frequently has only a minor interest in knowing whether failure on the part of the pupil is due to lack of comprehension of the instructions or due to the

inability to do the test elements. His primary interest is to find out whether the child possesses sufficient intelligence to deal with the total situation. And therefore the measurer of general intelligence may be right in contending that instructions should be absolutely uniform for all ages. Otherwise the total situation would not remain constant.

But it is unwise to carry over to pedagogical measurement a theory which is inapplicable. When an educator gives a vocabulary test, he is, as a rule, primarily interested to know what the pupil's vocabulary is, and only incidentally interested to determine whether the pupil possesses sufficient general intelligence to understand the instructions or overcome the mechanical difficulties of the form of the test. If a teacher measures her pupils' ability to add, she wants to know how well her children can *add*. She is not then interested in knowing how well they can understand her directions, or read printed instructions. She wishes to reduce these irrelevancies to a minimum. Only in a test of reading ability is it perfectly legitimate to make the instructions an integral part of the test itself. Nor is this primary interest peculiar to education. Many psychological tests which are designed primarily to measure intelligence prefer to measure it by means of the test material rather than by the instructions.

If the above distinction is sound it is legitimate to construct different instructions according to the age and ability of the pupils, provided whatever instructions are used give in every grade an adequate understanding of what is to be done, which means that if sixth-grade instructions are more difficult than third-grade instructions, the former must still be easy enough for each sixth-grade pupil to understand what he is to do. In essence this means that in pedagogical measurement adaptation has priority over uniformity. My thesis required both adaptation and uniformity because I think it is possible to secure both at once.

But it is frequently contended that there is no possibility of securing adequate adaptation together with uniformity.

It is claimed that the two characteristics are mutually antagonistic and that we cannot have our cake and eat it too. In recognition of this claim, Trabue in his Language Scales uses additional instructions and practice material for the younger children.

It is held by some that words which are appropriate for third-grade pupils would insult eighth-grade pupils and words appropriate for eighth-grade pupils would be beyond the comprehension of younger pupils. It may easily be doubted that third-grade children appreciate "baby talk" as much as is claimed. Nor is it impossible to find words sufficiently simple that younger pupils will understand them and at the same time so dignified that older pupils will not resent them.

When a test is being standardized for wide use throughout the country special care should be taken to see that instructions can really be kept uniform and yet be universally adequate and universally just. In the first place instructions should not require for their proper presentation material which some places may not have. Instructions should not require, for example, a blackboard unless a blackboard is likely to be available wherever the test is to be given. Again, the instructions should employ neither words nor illustrations which have local significance only. When Woody could not find a universal term in use which meant an *addition* example, he secured universality by giving other terms in common use and suggested that examiners use the terms current in the grade or locality where the testing is being done. Again, an examiner once discovered that the standard instructions lacked sufficient universality because they failed to take into consideration the fact that some pupils are left-handed. Illustration of elements conditioning universality could be multiplied.

4. ***The Order of Instructions Should Be the Order of Doing.***—It is probable that pupils can carry out instructions with greater ease when the order of the instructions is the order of doing. Long instructions are far more

tolerable when the steps in the description come in the same order as the steps of the process the pupil must go through. The demonstration is easier to imitate when the pupil does not find it necessary to transpose, in the process of doing the test, the steps observed in the demonstration. With our present meager knowledge and experience it may be unwise to hold to this principle absolutely and invariably, but there can be little doubt as to its general applicability. The following instructions for an aviation test which was still in the research stage when the armistice was signed, will illustrate this order of doing:

INSTRUCTIONS FOR AVIATION OBSERVERS' TEST

1. Seat the individuals to be tested in a well-lighted room and provide each with two pencils. Present the Observation Practice Chart directly in front of the subjects, on a level with, and fifteen feet from their eyes.
2. Distribute an Observation Test Sheet to each candidate, and say: *Write your name at the top of the sheet.*
3. When this is done, say:

   Imagine you are an army observer, and that the accuracy of the artillery fire will depend upon the accuracy of your observations. Among other things your test sheet will ask you to locate certain points, and to give the direction of certain points from certain other points. Watch while I show you how this is done, and how you are to record your answers.

   Look at question 1 on your test sheet. Suppose you were asked to locate J. You will note that according to the scale in the margin the center of J is 28 points to the right, and 4 points down. So the 28 would be written under the word "Right" and the 4 under the word "Down." Look at question 2. Suppose you were asked to give the direction of 9 from J. Imagine J to be the center of a circle which passes through 9. (Illustrate.) *Consider that directly above J is 0 degrees, directly east of J is 90°, directly south of J is 180°, directly west of J is 270°, and so on back to 0°. It is clear that the center of 9 is somewhere between 180° and 270°. It is in fact at 244°. Remember that 0° is always at the north and that the degrees increase clockwise. Do not make the mistake of recording the degrees counter-clockwise. The other questions need no explanation.*

   You will have 30 minutes, which allows less than a minute for each question. If you finish before time is called go back and try to improve your estimates.
4. Replace the Practice Chart with the Observation Test Chart; set the watch 9 hrs., 0 min., and 0 sec., and say *Begin!*
5. At 9 hrs., 4 min., 0 sec., say: *Even though you have not finished question 10, begin now on question 11.*
6. At 9 hrs., 11 min., 0 sec., say: *Even though you have not finished question 18, begin now on question 19.*

7. At 9 hrs., 20 min., 0 sec., say: *Even though you have not finished question 27, begin now on question 28.*
8. At 9 hrs., 30 min., 0 sec., say: *Stop! Pencils down!* and quickly collect the test sheets.
9. Immediately after collecting the papers, say: *Remember the importance of complete and accurate observations. You will have one minute to further study the chart, after which it will be covered, and you will be asked to compare what you have seen with what you see on another similar chart.*
10. Set the watch at 9 hrs., 0 min., 0 sec., and say *Begin!*
11. At 9 hrs., 1 min., 0 sec., remove the Observation Chart, distribute Memory Comparison Test sheets, and say: *Write your name at the top of each sheet.*
12. When names are written say: *Follow the instructions at the top of the test sheet. You will have twenty minutes. If you finish before time is called go back and try to improve your judgments. Compare what you remember of the chart you have been studying with this new one.*
13. At 9 hrs., 5 min., 0 sec., present the new chart and say: *Begin!*
14. At 9 hrs., 25 min., 0 sec., say *Stop! Pencils down!* and quickly collect papers.

5. ***Instructions Should Be Broken into Action Units.***—The strain upon the pupil's memory is not nearly so great when the instructions are broken into action units. Wherever possible the pupil should carry out directions before any other directions are given. The set of instructions just given illustrates instructions which are broken into action units. The instructions which follow are not broken into such units.

The experimenter holds the sheet before the class and says: "This sheet contains some incomplete sentences, which form a scale. This scale is to measure how carefully and rapidly you can think and especially how good you are in your language work.

"You are to write one word on each blank, in each case selecting the word which makes the most sensible statement.

"You may have thirty minutes in which to sign your name at the top of the page and write the words that are missing. The papers will be passed to you face downward. Do not turn them over until we are all ready. After the signal is given to start, remember that you are to write just *one* word on each blank and that your score depends on the number of perfect sentences you have at the end of thirty minutes."

It is easy to imagine just how little a pupil would remember of the key points in the latter set of instructions after the excitement of passing papers, writing names and the like. When the order of instructions is the order of doing, and when the instructions are properly segmented by action, the

instructions intimately concerned with each step of what the pupil is to do immediately precede that step. The pupil can give his undivided attention to that particular bit of instruction. When this principle is not satisfied the pupil is trying to grasp what is coming next and at the same time trying frantically to hold on lest what he has already heard escapes.

6. ***Instructions Should Equalize Interest.***—There are numerous factors besides interest which condition ability. Interest is dignified with special consideration because of its large effect upon the pupil's score. Interest determines effort. A pupil with high ability may show a range of interest from zero to high intensity, and hence a similar range of effort.

Shall standardization be upon a high plane of interest or upon a low plane? And how shall the desired stratification be secured? Experimental results have not yet shown whether it is easier to equate interest on a low plane, medium plane or high plane. Hence general common-sense experience must decide. Practical considerations rule out the offering of rewards high enough to secure the intensest possible interest. Normal life interests vary so greatly that they cannot be taken as a criterion. The fact that tests are not so educative when taken with low interest as when taken with high interest tends to rule out attempting an equalization of interest on a low plane. Furthermore, performance on one test does not seem to agree so well with performance upon a duplicate test when interest is on a low plane. In the absence of reliable evidence, the best guess is that performance is more constant and is a better index of the ability being measured when interest is at the maximum attainable by practicable methods.

What motivation can be legitimately employed? Unless such will defeat the object of the test, the pupil should be informed of the general purpose of the test and when it is not perfectly obvious, of the general method by which he is to be scored. A pupil will be more interested who is told

that the purpose of the test is to discover how rapidly he can read and then how accurately he can answer from memory questions upon what he has read, and hence his score will depend upon the number of seconds required to read a passage and the number of questions he can answer correctly upon what he has read. A detailed discussion of the purposes and methods of the test should not be attempted because of the necessity for brevity, and sometimes because of a necessity for concealing from the child the exact method of scoring. Secrecy is occasionally necessary in experimental work and in cases where the score is at the mercy of the pupil's honesty or lack of honesty.

The behavior of the child and the testimony of adults bear eloquent witness to the potency of rivalry as a begetter of interest. Probably no stimulus at the disposal of the school is so powerful, natural and generally healthful.

It is scarcely necessary to point out, however, that it will soon become impossible to secure interest through any method unless the pupils have an opportunity to learn how well they did on the test.

Project testing offers another method of not only securing interest but of securing reality and naturalness as well, for here the test itself becomes a motive. In a recent study of the Hope Farm School, N. Y., C. C. Certain used the project method of testing composition and penmanship. The Hope Farm pupils were requested to write a letter to the Horace Mann School pupils. Certain had made arrangements for the delivery of these letters and for returning replies from the Horace Mann pupils. Each Hope Farm pupil was provided with paper, envelope and the name of a boy and girl in the Horace Mann. Project testing is excellent where it is possible to give the time and take the trouble necessary to make it a success.

All the implications of the preceding paragraph have been that the device of securing interest by means of some form of rivalry is more artificial than project testing. We cannot be sure of this. Most of the games voluntarily selected by

children and adults would never be selected for their own sake. With children as well as adults rivalry is itself a project and is intrinsically satisfying. Remove the contest feature and how long would men and women lay card on card, or men punch ivory balls into holes with a long slender stick, or would war even remain the engaging pursuit that it is and the greatest game of all? Interest through projects is excellent, but interest through rivalry is not always artificial.

7. ***Instructions to Pupils Should Be Accompanied by Instructions to Examiners.***—Instructions to pupils should be accompanied by instructions to examiners telling how the test is to be applied, because it is a question which needs instructions more, pupil or examiner. Instructions to the examiner should be in steps easy to comprehend and follow. This easy use can be facilitated in two ways. First, the author of the instructions should formulate for the examiner the exact words to say to the pupils and insert between various units of directions to pupils, the necessary directions to the examiner. And, secondly, when the instructions to the examiner are inserted among instructions to pupils, the latter should be set off from the former in some convenient fashion. This can be done by numbering, paragraphing, underscoring, or italicizing the words to be said to the pupils. The sample set of broken instructions, given a few pages back, illustrates both the exact wording for pupils and a method of separating instructions to pupils from instructions to examiners.

CHAPTER IX

SCALING THE TEST

I. Percentile Scale—Percentile Unit

Why Scale Tests?—The fundamental aim of all testing is to reveal correct differences between pupils or groups of pupils. To reveal correct differences, a test must not only be valid but must possess, among others, the following characteristics. (Crude or exact scaling is prerequisite to each of the following traits.)

1. Every pupil should make some score larger than zero. If every pupil makes a zero score it is utterly impossible to tell which is the best, average, and stupidest pupil. If only one pupil out of a class makes zero, there is no way to determine just how much more stupid he is than the rest of the pupils. Zero-score pupils are unmeasured. The range of ability in a class is usually very great so, if the least able pupils are to make a score, the first elements of all difficulty tests must be within the ability of the least able and hence far easier than would be required for the abler pupils. The criterion requires that rate tests also must be composed of test elements whose difficulty is within the ability of all the pupils and which give a sufficiently long time limit. In an initial test in a recent Ph.D. research several pupils in one of the experimental groups made zero scores. In the final test, some time after, they made scores above zero. The conditions of the research required that the amount of their improvement be known. How much did the pupils improve? Nobody knows. It may have been and probably was a small amount, but it may have been enormous.

2. No pupil should make a perfect score. Perfect-score pupils are unmeasured just as zero-pupils are unmeasured. In the case of perfect scores it is not known how much better the pupils are and in the case of zero scores it is not known how much worse they are.

3. There should be no undistributed scores, whatever. A test often yields undistributed scores when there is not a single zero or perfect score, and these may occur anywhere between the lowest and highest scores inclusive. These undistributed scores are produced by coarse scoring. The coarsest possible method of scoring is the "all or none" method. To score pupils on a test as either "passed" or "failed" is an example of the "all or none" method, and gives very undistributed scores, for so far as the scores indicate all who receive a pass are exactly alike.

How fine should the scoring for a test be? The fineness of the scoring depends upon the uses to be made of the results. The following, however, will serve as a rough general rule: *Construct tests which will separate the pupils into at least seven groups of ability, and not less than thirteen if the data are to be used for correlation.* The above numbers, seven and thirteen, are minimum numbers. The finer the grouping the better. If the pupils are separated into less than seven groups of ability the results will have very limited uses, and if less than thirteen the influence of coarse scoring upon the coefficient of correlation will not be negligible. Among difficulty tests that one provides best against any sort of undistributed scores where the first test element is easy enough for all the pupils and where each succeeding element progressively increases in difficulty by small steps to a point beyond the ability of the ablest pupil. A very fine scoring of a few test elements will, however, produce the same effect as increasing the number of test elements. Stone, for example, has but twelve problems in his Reasoning Test in Arithmetic, yet it is possible to separate pupils into more than twelve groups by means of his test by giving credit for only parts of problems correct.

Each can easily figure out for himself just how to avoid undistributed scores in rate tests.

In the foregoing discussion of undistributed scores it has been assumed that each examiner will desire a score for each pupil, for unless such scores are secured a test cannot serve its most vital functions. In case only a class score is desired a few undistributed zero and perfect scores would do little or no harm if the median of pupil scores or the per cent of pupils doing a certain test element be the method of computing class scores. If, on the other hand, the mean of pupil scores is taken as the class score undistributed extremes may seriously effect the size of the class score.

4. The test should be scaled and the standardized method of scoring should utilize these refinements of the scaling. The scaling is a little more useful if the scale distance from one test element to the next is exactly equal to the scale distance between *any* two adjoining elements, i. e., if the scale progresses by equal steps or units of difficulty. The exactness of the scaling conditions the exactness with which differences between pupils can be measured.

5. A corrollary of the preceding paragraph is that a test should yield a statistical result. All measurement in descriptive words should give place to mathematical statement. Supervisors, for example, frequently rate teachers without developing any statistical system of recording and combining their ratings. It is mainly in the realm of subjective estimates that non-statistical measuring occurs. Recently an experiment was undertaken to determine by means of standard tests just how accurately supervisors could estimate the efficiency of certain teaching methods. When the time came to compare test results with the judgment of the supervisors, no worthwhile computations could be made, for the supervisors had not kept any statistical records.

6. Finally, correct differences cannot be revealed unless the two scores yielded by each rate test be reducible to a common denominator. Consider this situation from the Courtis Addition Test. Pupil A makes a speed score of 10

and an accuracy score of 90% while Pupil B makes a speed score of 12 and an accuracy score of 75%. Which pupil has made the better showing? As long as speed or accuracy is left free to fluctuate up and down in a sort of see-saw manner, no satisfactory comparisons between scores can be made until a table has been prepared for transmitting all scores to a constant speed or constant accuracy. Such a proposed table would tell us what would have been the accuracy of Pupil B had he worked at a speed of 10 examples instead of 12. Courtis [1] has originated a formula whereby speed and accuracy scores on his Arithmetic Tests, Series B, can be converted into a single score.

Perhaps the quickest method of determining the accuracy equivalence of a given amount of speed would be to adjust the weighting assigned until there has been secured the highest obtainable self-correlation between scores from two applications of the same rate test to the same pupils, when the scores correlated represent a combination score for both speed and accuracy.

There are numerous methods of scaling tests. First, there is the *goal scale* used by Courtis in connection with the Courtis Supervisory Tests. Any pupil whose score on the test falls between, say, 20 and 25 words spelled correctly on a particular spelling test is considered to have attained an appropriate spelling goal and is scored 1000. Any pupil who falls between, say, 17 and 20 is scored 500 and so on down to zero. Second, there is the *frequency-of-occurrence scale.* In the Jones' Vocabulary Test the score a pupil receives for knowing a certain word depends upon the frequency of that word's appearance in ten primers. In similar manner the degree of an individual's emotional aberration, as determined by the Kent-Rosanoff Free Association Test, is measured by the rarity of the individual's responses to the test. Then there is the percentile scale, age scale, grade scale, product scale, and T-scale. The last five are the ones

[1] S. A. Courtis and E. L. Thorndike, "Correction Formulæ for Addition Tests"; *Teachers College Record,* January, 1920.

more commonly used. These five only will be discussed in detail.

How to Construct a Percentile Scale.—Table 17 shows in the first column the number of questions in the Thorndike-McCall Reading Scale, Form 1. In the second column is shown the number of eleven-year-old pupils answering correctly a given number of questions. In the third column is shown the percentile corresponding to a given number of questions correct in the first column. The first and third columns constitute a *percentile table* for the eleven-year-olds for this particular reading scale. In similar fashion a percentile table could be prepared for this scale for any age or for any scale for any age, or, if desired, for this or any scale for various ages combined.

How was this percentile table constructed? In a later chapter will be found a detailed description of the computation of Q_1 (25 percentile), Q_2 (50 percentile or median), and Q_3 (75 percentile). All other percentiles are computed in similar manner. When scores are arranged in order of size and when counting is done in the direction of low to high, the 25 percentile is that score which is found by counting through one-fourth of the scores. The 50 percentile is found by counting through one-half of the scores. The 75 percentile is found by counting through three-fourths of the scores. Similarly the 10 percentile is found by counting through one-tenth of the scores, the 20 percentile by counting through two-tenths of the scores, the 35 percentile by counting through three-and-one-half tenths of the scores, and similarly for any other percentile. The lowest and highest scores may be considered the zero and 100th percentiles respectively.

Such is the procedure for counting percentiles directly. Percentiles may also be computed through the use of a table. When computed in the latter fashion, the percentiles do not so much represent what was actually found but what presumably would have been found had a very, very much larger number of pupils been tested.

TABLE 17

Shows the Construction of a Percentile Table for Eleven-Year-Old Children for the Thorndike-McCall Reading Scale, Form 1

Questions Correct	Number of Pupils	Percentile
0	1	0
1	1	
2	1	
3	1	
4	1	
5	1	
6	1	
7	5	
8	4	
9	2	
10	6	
11	4	
12	3	
13	4	
14	12	10
15	15	
16	22	
17	31	20
18	20	30
19	32	
20	42	40
21	35	50
22	40	60
23	32	70
24	29	80
25	22	
26	16	
27	16	90
28	13	
29	3	
30	4	
31	6	
32	0	
33	1	100

How to Interpret Percentile Scores.—Table 18 shows how to use percentile tables to interpret the scores of an

eleven-year-old pupil. Table 18 gives a percentile table for eleven-year-old pupils for each of three different tests. Test A is the Thorndike-McCall Reading Scale, Form 1, and the percentiles are taken from Table 17. Test B is an arithmetic test. Test C is a spelling test. The pupil made a score of 22 on Test A which is equivalent to a percentile of 60. He made a score of 39 on Test B which gives him an arithmetic percentile of 55, because a score of 39 is half way between 38 and 40 whose percentiles are 50 and 60. His spelling percentile is as shown, 50. Since a score of ten corresponds to three different percentiles, 40, 50, and 60, it is best to take the middle one. Had the pupil been ten years old instead of eleven, the percentile tables for ten-year-olds should be used in place of the percentile tables for eleven-year-olds.

TABLE 18

Shows How to Interpret the Scores on Three Tests for an Eleven-Year-Old Pupil by Means of Percentiles

TEST	PERCENTILE TABLES											PUPIL'S SCORES	PUPIL'S PERCENTILES
	0	10	20	30	40	50	60	70	80	90	100		
Test A	0	14	17	18	20	21	22	23	24	27	33	22	60
Test B	1	25	30	34	36	38	40	43	46	50	62	39	55
Test C	4	6	7	9	10	10	10	11	13	17	24	10	50
MEDIAN													55

The pupil's true percentile position on the three tests may, for most practical purposes, be taken as the median of his three percentiles, namely, 55. This percentile shows him to be slightly above the average for his age.

Theoretically, however, a pupil's mental index is not truly represented, when two or more tests are used, by the median or mean of his various percentiles on these tests. Pintner [2] used the percentile-scale method for his mental survey tests. In an admirable discussion of percentiles he points out that an additional step is necessary. Due to the fact that in-

[2] Rudolf Pintner, *The Mental Survey;* D. Appleton & Co., New York, 1918.

creased adequacy of measurement decreases variability it is necessary to have a super-percentile table which shows the true percentile value of various median percentiles. This super-percentile table is constructed just like a percentile table for a single test. The median percentiles for a large number of pupils of a given age take the place in the first column of Table 17 of the number of questions correct.

II. Age Scale—Growth Unit

How to Construct an Age Scale.—The construction of an age scale merely requires the determination of satisfactory age norms. The percentile units tell just how a pupil compares with all the other pupils of like age, if age is the basis, or like grade, if grade is the basis. Instead of interpreting a score for a ten-year-old pupil as the 25, 50 or 75 percentile of ten-year-olds, we could say that this ten-year-old pupil has a score equal to the average score for eleven-year-olds, or twelve-year-olds or any age above or below the age of the pupil in question. We could score the pupil 8, 10, 11.5, etc., meaning that he was respectively the equal of average eight-year-olds, average ten-year-olds, or half way between average eleven-year-olds and average twelve-year-olds. To do this would be to use what I have called the *growth unit*. Given a norm for each age, any pupil's test score may be readily transmuted into educational age and Educational Quotient (E.Q.) if the test is an educational test, or into mental age and Intelligence Quotient if the test is an intelligence test. Since the process is identical in both cases the transmutation is illustrated in Table 19 only for some educational tests.

How to Interpret Age Scores.—Table 19 is interpreted viz.: On Test A the average score is 4 for 8-year-olds, 8 for 9-year-olds and so on. The pupil in question made a score of 12. Since 12 is exactly the average score for 10-year-olds, the pupil's age score is 10. And since the pupil's chronological age is also 10 years, his Educational

TABLE 19

Shows the Use of Growth Unit and Educational Quotient (E.Q.) for Interpreting Educational Test Scores of a 10-Year-Old Pupil

Test	Age						Pupil's Test Score	Pupil's Age Score	Pupil's E. Q.
	8	9	10	11	12	13			
A — Av. Score	4	8	12	15	18	20	12	10	100
B — Av. Score	20	30	38	45	50	55	46	11.2	112
C — Av. Score	85	84	82	80	78	75	83	9.5	95
Median								10.	100

Quotient (E.Q.) is 100, computed thus: [(10 ÷ 10) × (100)]. An E.Q. of 100 means that the pupil is at age or exactly normal in the ability measured by Test A. The pupil's test score on Test B is 46, which is one-fifth or .2 of the distance from norms 45 and 50. Consequently the pupil's age score is 11.2, and his E.Q. is 112. Test C is scored in time units, hence the larger the score is, the less the ability is. The pupil is one-half or .5 of the way between average scores 84 and 82 which are the norms for ages 9 and 10 respectively. Hence the pupil's age score is 9.5 and his E.Q. is 95. The medians 10 and 100 tell us that in the median of all tests the pupil is a normal individual. Were this pupil 10 years and 6 months old chronologically, or 10.5 years old, his E.Q. for Test A would be 95.2 computed thus: [(10 ÷ 10.5) × (100)], his E.Q. for Test B would be 106.6 computed thus: [(11.2 ÷ 10.5) × (100)], his E.Q. for Test C would be 90.5 computed thus: [(9.5 ÷ 10.5) × (100)]. Sometimes it will be more convenient to reduce the age in years and decimals of years to months before computing E.Q. Thus the last E.Q. of 90.5 may be computed thus: 9.5 yrs. = 114 mos., 10.5 yrs. = 126 mos. [(114 ÷ 126) × (100)] = 90.5. The result is the same either way.

There are certain age scales which do not require the transmutation illustrated in Table 19. The Stanford Re-

vision of the Binet-Simon Intelligence Scale is an illustration of such a scale; whereas the Buckingham-Monroe Illinois Intelligence Examination is an illustration of one which does require transmutation. The Stanford Revision of the Binet-Simon Scale is so constructed that two months of age is allowed for every test element done correctly. Hence the original score comes out as the age score.

III. Grade Scale—Grade Variability Unit

Derivation of Grade Scale.—The scoring unit for time is the hour, for wealth is the dollar, for weight is the pound, for temperature is the degree, for distance is the foot and so on. The fundamental and most universal scoring unit for measuring educational achievement is the *P.E.*, *S.D.* or some other measure of the variability of pupil performance. The nature of this scoring unit is discussed in a later chapter. At this point it will be necessary only to consider how this unit is utilized in scale construction. The technique of grade scale construction is described in detail in Woody's *The Measurement of Some Achievements in Arithmetic*, Bureau of Publication, Teachers College, N. Y. C. A brief summary of the steps involved in this method of scale construction will be sufficient for our purpose. To date, this method has been applied to grades rather than ages.

Suppose an examiner wishes to make an addition scale for Grade III. The steps are, viz.:

1. He uses his judgment to select a large number of addition examples, gradually varying in difficulty from the easiest possible example to very difficult examples.

2. He tries out these examples upon a large number of chosen-at-random, third-grade pupils.

3. He computes the per cent of pupils correctly solving each example. If 90% solve a given example, obviously this example is very easy and hence has a low scale value. If 50% solve a given example, that example is of average diffi-

culty and occupies a medium high position on the scale. If only 5% solve a given example, this example is very difficult and occupies a high position on the scale. Thus an example's position on the addition scale is determined by the per cent of pupils correctly solving it. The larger the per cent is, the less the difficulty, and the lower the example's position on the difficulty scale.

4. By means of Table 20, he converts these per cents into *P.E.* units of difficulty or *P.E.* distances above and below the third-grade median.

TABLE 20

Showing the *P.E.* Values above or below the Median Corresponding to the Per Cent of Pupils Correctly Doing a Given Test Element. Before Reading, Subtract the Per Cent Correct from 50% Examples: 83.60% did test element A. 50% — 83.60% = — 33.60% = 1.45 *P.E.* Below Median. 2.15% did test element B. 50% — 2.15% = 47.85% = 3.0 *P.E.* above the median.

x / P.E.	.00	.05	x / P.E.	.00	.05	x / P.E.	.00	.05	x / P.E.	.00	.05
0	0000	0135	1.5	3441	3521	3.0	4785	4802	4.5	4988	4989
.1	0269	0403	1.6	3597	3671	3.1	4817	4831	4.6	4990	4991
.2	0536	0670	1.7	3742	3811	3.2	4845	4858	4.7	4992	4993
.3	0802	0933	1.8	3896	3939	3.3	4870	4881	4.8	4994	4994.6
.4	1063	1193	1.9	4000	4057	3.4	4891	4900	4.9	4995.2	4995.7
.5	1321	1447	2.0	4113	4166	3.5	4909	4917	5.0	4996.2	4996.6
.6	1571	1695	2.1	4217	4265	3.6	4924	4931	5.1	4997.1	4997.4
.7	1816	1935	2.2	4311	4354	3.7	4937	4943	5.2	4997.7	4998.0
.8	2053	2168	2.3	4396	4435	3.8	4948	4953	5.3	4998.2	4998.4
.9	2291	2392	2.4	4472	4508	3.9	4957	4961	5.4	4998.6	4998.8
1.0	2500	2606	2.5	4541	4573	4.0	4965	4968	5.5	4999.0	4999.1
1.1	2709	2810	2.6	4602	4631	4.1	4971	4974	5.6	4999.2	4999.3
1.2	2908	3004	2.7	4657	4682	4.2	4977	4979	5.7	4999.4	4999.5
1.3	3097	3188	2.8	4705	4727	4.3	4981	4983	5.8	4999.55	4999.6
1.4	3275	3360	2.9	4748	4767	4.4	4985	4987	5.9	4999.65	4999.7

Suppose the per cent of pupils working correctly examples A through G were as in Table 21. Table 21 shows how to use Table 20 for converting per cents correct into *P.E.* distances above and below the median of Grade III.

TABLE 21

Showing the Use of Table 20 for Converting Per Cents Correct Into *P.E.* Distances from the Median

Examples	A	B	C	D	E	F	G
Per cent correct........	95.	85.	60.	50.	40.	12.2	2.2
Subtract from 50 per cent	—45.	—35.	—10.	00.	10.	37.8	47.8
P.E. Distances...........	— 2.45	— 1.55	— 0.4	00.	0.4	1.75	3.0

The difference in difficulty between Example A and Example B is (95 — 85) or 10% which is (2.45 — 1.55) or .90 *P.E.* The difference between C and D is also 10%, but the *P.E.* difference is only .4 or less than half the difference between A and B. The difference between D and E is also 10% and the *P.E.* difference is .4. The difference between F and G is likewise 10% and the *P.E.* difference is 1.25. Obviously, differences in per cent tell us almost nothing about the differences in difficulty. The real difference in difficulty is shown not by the per cents but by the *P.E.* values. The C D and D E differences are equal both in per cents and *P.E.'s*. This is because D is at the median while C and E are at like positions below and above D. The F G difference is the largest because these are the most extreme per cents. Example A has the largest per cent correct and yet it is farthest below the median. This is because Table 21 shows a scale of difficulty. That example which the largest per cent of pupils work correctly is the easiest example.

5. He determines the *P.E.* distance of the zero point of addition ability from the third-grade median. In locating the zero point, the point of view changes. We no longer enquire what per cent of pupils worked any one example but what per cent of the third-grade pupils did not succeed in working a single example. Suppose that 4% failed to work a single example. According to Table 20 this is 2.6 *P.E.* below the third-grade median and is the zero point.

6. He determines how many *P.E.* each example is above the zero point, and the scale is finished. The *P.E.* value of each example above zero is shown below, and was computed by algebraically subtracting the distance of the zero point below the median, namely, — 2.6 *P.E.,* from the *P.E.* value of each example as shown in Table 21.

Examples	A	B	C	D	E	F	G
P.E. Value above Zero...	.15	1.05	2.2	2.6	3.0	4.35	5.6

7. Sometimes he eliminates from the scale those examples which do not fall at equal *P.E.* intervals. Woody's

Arithmetic Scales, Series B, represent such a selection from his longer Series A and gives a scale progressing by approximately equal steps.

8. If the addition scale is being constructed for the entire elementary school instead of for only one grade, he repeats steps 2, 3 and 4 for each elementary school grade as has just been done for Grade III.

9. He computes the *P.E.* distance from each grade median to the adjoining grade median or medians. This is done by computing the per cent of pupils in one grade having scores which are larger than the median of the adjoining grade. This per cent, when read in Table 20, shows the *P.E.* distance between the medians of the two grades. The two possible direct and several possible indirect determinations are averaged to get a truer *P.E.* distance.

10. By the use of these intervals, he converts the *P.E.* distance of each example from its own grade median into *P.E.* distances above the common zero point of reference.

11. He then determines the final elementary-school *P.E.* value of each example and hence its position on the scale by averaging its *P.E.* values in the different grades. In making this average, the values in certain grades are usually given more weight than the values in other grades, for the reason that *P.E.* values which fall near the middle of a grade distribution are more reliably determined than when they fall near the extremes. The method of weighting is described in Woody's *The Measurement of Some Achievements in Arithmetic.*

12. When an elementary-school scale alone is desired much labor can be saved by computing from the outset the per cent of pupils in all the grades combined who solve each example correctly. In this way only one *P.E.* value of each example is read from Table 20. This value is referred to a zero determined from the per cent of all the pupils who make no score, rather than from the per cent of the second or third grade, and the scale is finished.

This short method assumes that ability in Grade III

through VIII combined is distributed according to the normal frequency surface. This is not a specially violent assumption. The curve of ability for all these grades combined would, however, be slightly flat-topped, especially if Grades II and III were included.

13. Still another short method would be to use the scale determined for the sixth grade as the elementary school scale. Standing as it does about mid-way between the fourth and eighth grade, results from it could be considered typical of the whole school.

Constancy of the P.E. in Grade Scales.—If we were to measure carefully the length of a bar of iron in inches and were to perfectly preserve the bar and measure it again in 200 years, the bar would still be the same number of inches in length. The fifth example on Woody's Addition Scale, Series B, has a *P.E.* value of 3.26. Will this example, $3 + 1 =$, have a value of 3.26 *P.E.* 200 years from now?

There are at least two forces operating to cause a fluctuation from time to time in the value of an example, or any other test element for that matter. There are, first, changes in courses of study, efficiency of instruction and the like. Probably the chief difficulty in the above example was the plus sign. The pupils found it easier to add 72 and 26 when the 26 was written under the 72 than to add $3 + 1$. The very fact that this simple example proved so difficult will undoubtedly attract the notice of teachers and a special effort will be made to familiarize pupils with signs early in their school career. Courses of study may be altered to bring instruction in the four algebraic signs earlier than at present. The result would be a rearrangement of the problems in the scale. Any change that was partial or otherwise to a particular test element would upset its present P.E. value.

A second factor is an alteration in the classification of pupils. The use of tests is gradually improving the accuracy of classification, which in turn is reducing the amount of overlapping of ability between grades, which in turn is in-

creasing the interval between grades and influencing the form of distribution, all of which is bound to influence the *P.E.* values of various test elements. Again, the grade, always a rather artificial feature, is becoming more or less meaningless. Some schools have reduced the eight grades to seven, others have eliminated grades by a departmentalization, others have crossed grade lines with section classification on the basis of educational and mental tests. It is becoming increasingly probable that future test constructors will prefer age to grade as a basis for scale construction.

III. Product Scale—Variability-of-Adult-Performance Unit

Ayres' Handwriting Scale.—So far as I recall, Ayres' [3] Handwriting Scale is the only scale which makes use of this variability-of-adult-performance unit for scoring the achievement of pupils. Most handwriting scales determine values on the scale by the judgment of adults rather than by the achievement or performance of adults. A quotation from Ayres will bring out very clearly the relationship between his scale and Thorndike's scale and will show his reasons for adopting a different scoring unit.

"The Thorndike Scale

"This (Ayres' Scale) is not a pioneer piece of work in this field, although it is different in method from anything of the sort previously attempted. The credit of developing the first measuring scale for handwriting belongs to Professor Edward L. Thorndike of Teachers College, Columbia University. The publication, in March, 1910, of his handwriting scale constituted a most important contribution not only to experimental pedagogy but to the entire movement for the scientific study of education.

"As Professor Thorndike says in his introduction, previ-

[3] Leonard P. Ayres, *A Scale for Measuring the Quality of Handwriting of School Children, No. 113;* Russell Sage Foundation, N. Y. C.

ous to that time educators were in the same condition with respect to handwriting as were students of temperature before the discovery of the thermometer. Just as it was then impossible to measure temperature beyond the very hot, warm, cool, etc., of subjective opinion, so it had been impossible to estimate the quality of handwriting except by such vague standards as one's personal opinion that given samples were very bad, bad, good, very good, etc. Professor Thorndike's scale for the handwriting of children is based on the average or median judgments of some 23 to 55 judges who graded samples of writing into groups by what they considered equal progressive steps in general merit.

"Legibility as a Criterion

"The method by which the present scale has been produced, and the criterion on which it rests as a basis, differ radically from those adopted by Professor Thorndike. The difference in the bases is that in the present case legibility has been adopted as a criterion for rating the different samples in place of 'general merit' used as the basis of Thorndike's scale. This change substitutes function for appearance as a criterion for judging handwriting.

"There are two arguments for adopting the new criterion. In the first place the prime importance of writing is to be read, and hence it has seemed worth while to adopt 'readability' as the basis for rating samples of handwriting. In the second place legibility possesses the advantage of being measurable in definite quantitative units through finding the amount of time required to read with a given degree of accuracy, a given amount of matter in the handwriting being studied. The criterion of general merit is not susceptible of any such evaluation.

"The method whereby the new scale has been produced differs from the method employed in producing the previous scale in that it is based on the distribution of the recorded time required on the average by a number of readers to read the samples of writing, rather than on the average of their judgments concerning what they considered equal steps in general merit."

Having, with proper experimental precautions, determined the average time required by 10 individuals to read each of 1578 specimens of handwriting collected from the upper grades of the elementary school, he plotted all these rates into a frequency surface. The base line of the frequency surface was divided into ten equal intervals, the lowest rate being called 0, the highest rate being called 100, and the mean rate on all specimens being labeled 50. The samples of handwriting having rates corresponding to the points 20, 30, . . . 100, were located and printed as a scale.

The scale is used viz.: A pupil's handwriting is moved along the scale until a writing of the same quality is found. The pupil's specimen is scored 20, 30, 40, or above according to the value of the scale specimen with which it corresponds in quality.

The reader's attention is called to the following points and questions: First, legibility and not "general merit" is the criterion for this scale. Is legibility the only important element in the world's practical evaluation of handwriting? Second, a pupil's handwriting is scored by a judgment comparison with the scale. Does this final use of judgment make void the legibility criterion?

IV. Product Scale—Variability-of-Judgment Unit

Derivation of Product Scales.—Hillegas' English Composition Scale, Starch's Handwriting Scale and Thorndike's Drawing Scale are typical instances of pure product scales. All were constructed in the same manner, which was as follows:

1. The scale constructor selects many specimens of, say, composition which vary by small amounts from compositions of zero merit up to, say, the highest quality of composition produced by the best authors.

2. He asks many presumably competent judges to arrange the compositions in order of merit and also to designate the specimen which is, in their judgment, of just zero merit.

3. He computes from these rankings the per cent of judges who rated specimen A better than specimen B, better than specimen C and so on. Then he computes the per cent of judges who rated specimen B better than specimen C, better than specimen D and so on. He continues this process until he has a table showing the per cent of judges who rate each specimen better than every other specimen. The per cent of judges rating a very poor specimen better than a very good one is likely to be zero, while the per cent rating the specimen of high merit better than the specimen of low merit is likely to be 100. Per cents of better judgments will range all the way from zero to 100.

4. He subtracts 50 per cent from all the above per cents.

5. He determines the *P.E.* difference in merit between each specimen and every other specimen by looking up these remainder per cents in Table 20. Table 22 illustrates the process.

TABLE 22

Shows How to Use Table 20 to Convert Per Cent of Better Judgments into *P.E.* Differences in Merit Between Composition Specimens A, B, C, etc.

Specimens	A > T	N > A	B > N	K > B	E > K	L > E
Per cent	50	75	75	84.41	64.47	91.13
Per cent minus 50	00	25	25	34.41	14.47	41.13
P.E. Difference..	00	1.0	1.0	1.5	.55	2.00

6. He not only determines the *P.E.* difference AB, AC, AD, etc., and BC, BD, BE, etc., directly, but he determines these differences in many indirect ways as well. Thus, for example, the distance N A, above, equals T N minus T A, the distance B N equals A B minus A N, the distance L E equals [(T L) minus (T A + A N + N B + B K + K E)]. There are many other indirect ways of determining the *P.E.* difference between any two specimens.

7. The mean of all possible direct and indirect determinations of *P.E.* differences is computed to get the true difference. The greater the indirectness the less the

weight given to the determination in computing this mean *P.E.* difference.

8. He arranges the specimens in order of merit recording the *P.E.* distance each is above the preceding one, thus:

Specimen	T	A	N	B	K	E	L
P.E. Distance.......		0	1.0	1.0	1.5	.55	2.0

9. He records from the original data the number of judges indicating each specimen as of just zero merit. Some will indicate, say, K, some B, some N, some A, some T, and some specimens which are below A and T in merit.

10. He computes the median zero specimen. Let us suppose that the median specimen is found to be A.

11. He computes the *P.E.* distance each specimen is above the zero specimen and calls this its scale value. Since A and T are of equal merit the scale becomes as follows:

Specimen	A or T	N	B	K	E	L
Scale Value....	0	1.0	2.0	3.5	4.05	6.05

12. Beginning with the zero specimen he selects others above it such that distances between specimens will be about 1 *P.E.* Smaller scale steps are probably not desirable for scales which are to be widely used because a difference of 1 *P.E.* is a difference which only 75 out of 100 judges can see. Smaller-step scales may be valuable for scientific work, or for use by individuals who are specially expert in detecting subtle differences in merit. When two or more specimens have approximately the same scale value they may all be presented in order to give a wider range of composition type, or that one may be selected which shows the least disagreement among judges. The Thorndike Extension of the Hillegas Scale adopted the former method and the Nassau Extension of the Hillegas Scale the latter.

Validity and Constancy of Judgment Units.—What is the validity of this *P.E.* as a unit of measurement? Product scales were made possible by the formulation of the now famous Cattell-Fullerton theorem, and by the ingenious application by Thorndike of this theorem in the construction

of educational scales. Courtis has reported an experiment which was conducted to test the validity of this basic theorem; namely, *differences which are equally often noticed are equal unless they are always noticed or never noticed.* Courtis wanted to know whether differences which are equally often noticed really are equal. To test this he made a product scale of areas instead of compositions or specimens of handwriting. After determining the differences between areas of variously shaped figures by means of judgments, he determined the differences by actual measurement. The differences as determined by judgments followed the principle of Weber's law, i. e., when the area was small, a slight increase or decrease in area could be seen; when the area was large, a considerable change of area was necessary in order that judges might be able to notice the difference. In other words, equally often noticed differences were equal for areas of about the same size only. The theorem does not hold for widely separated areas. Does it hold for specimens of penmanship widely separated in merit? Presumably it does not, if there are absolute differences in merit of handwriting in the same sense that there are absolute differences in area.

Even if this last is true, we need not lose confidence in our product scales. Education is interested in many kinds of differences. It would be valuable to know them all. There are absolute differences such as Courtis points out. There are difficulty differences, and this is the kind of difference Woody's arithmetic scales bring out. Product scales measure judgment differences. The value on percentile, age and grade scales are determined by how difficult pupils actually find the test elements. These scales could be converted into product scales by determining the difficulty of each test element, not by the achievement of the pupils, but by the opinion of adults. This has not often been done simply because education is far more concerned with how difficult test elements actually are than how difficult some-

body thinks they are. But in the realm of composition, handwriting and the like, we are not primarily concerned with difficulty but with merit, and we are less concerned with an absolute merit than we are with the merit as determined by the opinion of competent judges, in the way that competent judges practically operate outside or inside the schools.

Is the judgment scoring unit constant? The meter was originally defined as one ten-millionth of the distance from the pole to the equator. Alteration of this distance through the centuries due to the contraction or expansion of the earth would, of course, alter the meter, especially if a redetermination became necessary because of the loss of the meter bar carefully preserved at Paris. Alteration of this distance due to the subjectivity of the determiner would also alter the meter. As a matter of fact no two determinations of the pole to equator distance have turned out to be exactly the same. Consequently the meter is now measured in terms of so many wave lengths of a certain radiation. What forces are operating to produce an inconstancy of *P.E.* in, let us say, a composition scale?

Only the two most likely forces need be mentioned. First, it is possible to discriminate finer shades of composition merit. There is certainly room for improvement in this respect. So far as most of us are concerned there is "low visibility" when it comes to evaluating composition merit. The effect of a more microscopic eye would be to make *P.E.* smaller than it is at present. Second, it is possible that future judges will have a different opinion from present-day judges as to what constitutes merit in a composition. It is conceivable, but scarcely probable, that a literary dictator will arise whose popularity will be so great as to completely change the current of the world's literary appreciation. The nibbling of literary radicals is undoubtedly producing small but continuous changes in the weight we attach to each of the numerous factors entering into a

composition. As I sit in Morningside Park, numerous small caterpillars are spinning a basket web in a nearby bush. A large red spider has gratefully attached one edge of his web to the caterpillar's basket. Many of the caterpillars are so interested in the mechanics of their work that they often imprison themselves in their own silken net. The spider doubles their imprisonment by rolling them into little round balls. The radicals claim that our meticulous teaching methods have in similar fashion imprisoned the "rare spirit's wings" in the mesh of mechanics. We are only just beginning to investigate the extent to which literary evaluations vary with the age and sex of the judge, the type of literary instruction he has had, the purpose for which the compositions were written and the like.

Peculiarity of Product Scales.—Composition, handwriting and drawing scales are peculiar in that they are not tests at all. They are scoring instruments. For this reason as well as for the manner of their construction they are called *product scales* to contrast them with percentile, age, and grade scales which together are usually called *performance scales*. The performance scales are placed in the hands of the pupils in order to test them. They are real testing instruments and not scoring instruments. Thorndike's Visual Vocabulary Scale A-2 is the test instrument. The scoring instrument is a correctly filled out test. Woody's Addition Scale is a test instrument. The scoring instrument is a set of correct answers. Collection of the pupils' composition specimens is the composition test. The composition scale is only the scoring instrument. In the case of performance scales the dramatic instrument is not the scoring instrument but the testing instrument. To date, the scoring instrument has been assumed to be satisfactory. In the case of product scales the situation is exactly reversed. As will be suggested later, both scoring instrument and what is scored may be scaled. The following table will make clearer the relation between what is scored, the scoring scale, the scoring instrument, and the scale unit.

Thing Scored	*Scoring Scale*	*Scoring Instrument*	*Scale Unit*
Man's height	Distance	Yard stick	Yd. ft. in.
Until train leaves	Time	Watch	Hr. min. sec.
Heat of water	Temperature	Thermometer	Degree
Courtis Arith., Series B			
(a) Speed	Speed	None	The example
(b) Accuracy	Accuracy	Correct answers	The example
Woody Arith., Series B	Difficulty	Correct answers	*P. E.*
Thorndike-McCall Reading Scale	Difficulty	Correct answers	T.
Starch Handwriting	Quality	Handwriting specimens	*P. E.*
Composition	Quality	Composition specimens	*P. E.*

CHAPTER X

SCALING THE TEST. T SCALE—AGE VARIABILITY UNIT

I. The Method of Scale Construction

Preparation of Test Material.—Recently I undertook, at the suggestion of Thorndike, the task of constructing a much-needed series of reading scales. A careful study of previous methods of scale construction led to the conviction that there was great need for revision. Perhaps the best way to show the method evolved and why it was evolved to correct some of the defects of existing methods is to describe in detail just how one of these reading scales was constructed. The steps in the process, with some alterations, were as follows:

1. Selections of prose and poetry were made which were brief, which gradually varied in difficulty from very easy to very difficult, which were fairly representative of reading material both in school and out, which were reasonably free from technical terms, and which were equally fair to rural and urban children.

2. Questions were formulated which could be answered from or inferred from or were related to the reading selections, which would yield brief, scorable answers, which were unambiguous, whose difficulty approximated that of the selection, and which were independent either in wording or difficulty of any preceding or succeeding questions, and which were numerous enough to make up one test of the desired length.

3. Several experienced teachers answered all the questions and assisted in arranging the questions in the order

of their difficulty for school children, i. e., the questions judged to be easiest were placed at the beginning of the test and the most difficult ones at the end.

4. The test when arranged was mimeographed along with the instructions to pupils. So far as possible the type and spacing and other features were identical with that to be used in the final printed scale.

Scaling the Individual Questions.—

5. The test was applied to a few hundred pupils in grades III through VIII. The total population in each school was tested so as to yield a rough approximation to a normal frequency distribution.

6. The pupils' answers to the questions were scored as either right or wrong.

7. The few questions which in the try-out proved ambiguous or unscorable or were otherwise unsatisfactory were eliminated.

8. The results for the remaining questions were tabulated by each question for each pupil. The following shows how the tabulation was made.

Pupil	*Selection I*	*Selection II*	*Selection III, etc.*
	1 2 3 4	1 2 3 4 5	1 2 3 "
S. A.	1 1 1 1	0 1 1 0 0	0 0 0
R. N.	1 0 1 1	1 1 1 1 1	0 1 0

9. The total number of pupils answering each question correctly was computed and divided by the number of pupils tested to get the per cent of correct answers.

10. This per cent was found in Table 23 and the corresponding difficulty or *S.D.* value was read. The difficulty value of each question so found was only roughly correct. Table 23 assumes a normal distribution of ability among the pupils. This assumption is sufficiently met by using only pupils in one grade or only pupils of one age. Combining grades makes the curve of ability somewhat too flat on the top. But the values from combined grades were accurate enough for the purpose. Computing from all grades combined saves much time.

How Table 23 was used to convert per cents correct into *S.D.* values, i. e., difficulty values, is shown below.

Selection	I			VIII			
Questions ..	1	2	3	1	2	3	4
Per cent correct	99.99	99.7	98.13	0.61	0.0003	0.008	0.53
S.D. Value ..	13.	22.	29.	75.	90.	88.	76.

TABLE 23

Shows the *S.D.* distance of a given per cent above zero. Each *S.D.* value is multiplied by 10 to eliminate decimals. The zero point is 5 *S.D.* below the mean

S.D. Value	Per cent	*S.D.* Value	Per cent	*S.D.* Value	Per cent	*S.D.* Value	Per cent
0	99.999971	25	99.38	50	50.00	75	0.62
0.5	99.999963	25.5	99.29	50.5	48.01	75.5	0.54
1	99.999952	26	99.18	51	46.02	76	0.47
1.5	99.999938	26.5	99.06	51.5	44.04	76.5	0.40
2	99.99992	27	98.93	52	42.07	77	0.35
2.5	99.99990	27.5	98.78	52.5	40.13	77.5	0.30
3	99.99987	28	98.61	53	38.21	78	0.26
3.5	99.99983	28.5	98.42	53.5	36.32	78.5	0.22
4	99.99979	29	98.21	54	34.46	79	0.19
4.5	99.99973	29.5	97.98	54.5	32.64	79.5	0.16
5	99.99966	30	97.72	55	30.85	80	0.13
5.5	99.99957	30.5	97.44	55.5	29.12	80.5	0.11
6	99.99946	31	97.13	56	27.43	81	0.097
6.5	99.99932	31.5	96.78	56.5	25.78	81.5	0.082
7	99.99915	32	96.41	57	24.20	82	0.069
7.5	99.9989	32.5	95.99	57.5	22.66	82.5	0.058
8	99.9987	33	95.54	58	21.19	83	0.048
8.5	99.9983	33.5	95.05	58.5	19.77	83.5	0.040
9	99.9979	34	94.52	59	18.41	84	0.034
9.5	99.9974	34.5	93.94	59.5	17.11	84.5	0.028
10	99.9968	35	93.32	60	15.87	85	0.023
10.5	99.9961	35.5	92.65	60.5	14.69	85.5	0.019
11	99.9952	36	91.92	61	13.57	86	0.016
11.5	99.9941	36.5	91.15	61.5	12.51	86.5	0.013
12	99.9928	37	90.32	62	11.51	87	0.011
12.5	99.9912	37.5	89.44	62.5	10.56	87.5	0.009
13	99.989	38	88.49	63	9.68	88	0.007
13.5	99.987	38.5	87.49	63.5	8.85	88.5	0.0059
14	99.984	39	86.43	64	8.08	89	0.0048
14.5	99.981	39.5	85.31	64.5	7.35	89.5	0.0039
15	99.977	40	84.13	65	6.68	90	0.0032
15.5	99.972	40.5	82.89	65.5	6.06	90.5	0.0026

S.D. Value	Per cent	S.D. Value	Per cent	S.D. Value	Per cent	S.D. Value	Per cent
16	99.966	41	81.59	66	5.48	91	0.0021
16.5	99.960	41.5	80.23	66.5	4.95	91.5	0.0017
17	99.952	42	78.81	67	4.46	92	0.0013
17.5	99.942	42.5	77.34	67.5	4.01	92.5	0.0011
18	99.931	43	75.80	68	3.59	93	0.0009
18.5	99.918	43.5	74.22	68.5	3.22	93.5	0.0007
19	99.903	44	72.57	69	2.87	94	0.0005
19.5	99.886	44.5	70.88	69.5	2.56	94.5	0.00043
20	99.865	45	69.15	70	2.28	95	0.00034
20.5	99.84	45.5	67.36	70.5	2.02	95.5	0.00027
21	99.81	46	65.54	71	1.79	96	0.00021
21.5	99.78	46.5	63.68	71.5	1.58	96.5	0.00017
22	99.74	47	61.79	72	1.39	97	0.00013
22.5	99.70	47.5	59.87	72.5	1.22	97.5	0.00010
23	99.65	48	57.93	73	1.07	98	0.00008
23.5	99.60	48.5	55.96	73.5	0.94	98.5	0.000062
24	99.53	49	53.98	74	0.82	99	0.000048
24.5	99.46	49.5	51.99	74.5	0.71	99.5	0.000037
						100	0.000029

Selection and Arrangement of Final Test.—

11. The questions were rearranged in order of actual difficulty. Any question which had been greatly misplaced originally was eliminated entirely. Scales as they have usually been scaled need to be scaled over again just as soon as they are finished. Because the original test has usually been much larger than the final test and because some of the test elements have been shifted, in the process of scale making, far from their original position, it has been found that the difficulty of test elements is different in the final scale from what their difficulty was in the original setting.

The difficulty of any question is partly a function of its real difficulty, partly a function of some preceding question which may give a right or wrong mental set, and partly a function of its distance from the first question. Place a question nearer the beginning of a test and it tends to become easier. Remove it from its surrounding questions and unless it is highly independent of them its difficulty tends to vary. In carefully constructing a scale it is advisable for

these reasons, to eliminate a question whose position must be greatly altered. For the same reasons it is well to go through this preliminary scaling in order that the final scaling will have permanence.

All transient questions might have been eliminated. Thus far the difficulty of each question has been determined for the total group only. This is usually all that is necessary. But it may be the case that the order of difficulty for certain questions would be markedly different for different grades. Questions whose difficulty varies from group to group in this fashion are called *transients*. Undoubtedly a scale would be improved by the elimination of the worst of these transients. This may be done by repeating steps 9 and 10 for each grade separately. It is doubtful whether the scale freed from transients is enough superior to justify the large amount of extra labor required.

12. Any serious gap of difficulty in the scale which could not be filled by shifting a question from one position to another was filled by combining two or more questions and treating them as a single question in the final scale. Just as it was possible to compute, in step 9, the per cent of pupils who answered correctly each question and to convert this per cent, in step 10, into a scale value, so it was possible to compute for this same group the per cent of pupils who answered any two or any three questions correctly and to convert this into a comparable scale value.

It should be noted that the per cent of pupils answering two questions correctly could not be found by averaging the per cent of pupils doing one of the questions correctly with the per cent answering the other one correctly. It should also be noted that two questions so combined were ever after treated as a single question. Both answers had to be correct before a pupil could receive credit for the question.

13. The material thus selected and arranged was printed along with instructions, and other advice, in its final booklet form.

Application and Scoring of Final Test.—

14. An effort was made to select for final testing purposes a group of schools which when combined would have at least 500 pupils between the ages of 12.0 and 13.0 and which when combined would give fairly representative pupils of all ages, i. e., the per cent of pupils of each level of ability found in any total unselected group. It was, of course, impossible to find in schools representative pupils who were very young or very old.

15. The test was applied to all pupils in grades III through VIII inclusive and to all pupils between the ages of 12.0 and 13.0 whether they were in ungraded classes or high school. When the test was given in cities so large that the total school population was not measured caution had to be exercised to test just the right number of 12-13 year old high-school pupils.

16. Each question was carefully scored as either right or wrong according to a set of guiding principles formulated for the purpose of making all scoring as uniform as possible. There is considerable evidence to indicate that the method of scoring with partial credits for questions in various stages of correctness is not enough superior to simply calling the pupil's answer right or wrong to justify the extra trouble. The method of right-or-wrong scoring is not however obligatory.

17. All correct answers to each question, the worst answers which were accepted and the best answers which were rejected as well as other sample answers were carefully tabulated as found to make a scoring key for the guidance of all future scoring both prior to the completion of the scale and after.

18. The test booklet for each pupil was thrown into the pile for his age and grade. The following illustrates how this classification was made. The Roman numerals represent grades and the Arabic numerals represent chronological ages. There was a stack of papers for each Roman numeral, except in cases where no pupils of a given age group were found in the grade indicated.

6—7	7—8	8—9	9—10	10—11	11—12
III	III	III	III	III	III
IV	IV	IV	IV	IV	IV
V	V	V	V	V	V
VI	VI	VI	VI	VI	VI
VII	VII	VII	VII	VII	VII
VIII	VIII	VIII	VIII	VIII	VIII

12—13	13—14	14—15	15—16	16—17	17—18
III	III	III	III	III	III
IV	IV	IV	IV	IV	IV
V	V	V	V	V	V
VI	VI	VI	VI	VI	VI
VII	VII	VII	VII	VII	VII
VIII	VIII	VIII	VIII	VIII	VIII

Scaling the Total Number of Questions Right.—

19. The total number of questions (single, double, or triple), which were answered correctly by each twelve-year-old pupil, was determined. This is shown in the first and second columns of Table 24.

20. The per cent of all twelve-year-old pupils who exceeded no questions plus half those who did no questions was computed. Thus of the 500 twelve-year-olds shown in Table 24, 497 made scores larger than no questions correct. Half of those answering no questions correctly plus 497 makes 498.5 which is 99.7 per cent of 500. In similar fashion there was computed the per cent of those exceeding one question plus half those doing one question, and then the per cent of those exceeding two questions plus half those answering two questions, and so on. This yields the per cents shown in the third column of Table 24.

21. These per cents were converted into *S.D.* values or scale scores by the use of Table 23. Thus a per cent of 99.7 is equivalent to a scale score of 23, and a per cent of 99.3 is equivalent to a scale score of 25, and so on for the other per cents. This means that pupils who are unable to answer a single question on the test should be assigned a scale score of 23, those who answer one question correctly should be assigned a scale score of 25 and so on.

TABLE 24

Shows How to Scale Total Scores

Total No. Questions Correct	No. of Twelve-Year-Old Pupils	Per Cent Exceeding Plus Half Those Reaching	Scale Score
0	3	99.7	23
1	1	99.3	25
2	2	99.0	27
3	1	98.7	28
4	2	98.4	29
5	2	98.0	29
6	2	97.6	30
7	2	97.2	31
8	4	96.6	32
9	2	96.0	32
10	2	95.6	33
11	10	94.4	34
12	3	93.1	35
13	8	92.0	36
14	8	90.4	37
15	13	88.3	38
16	15	85.5	39
17	18	82.2	41
18	28	77.6	42
19	26	72.2	44
20	34	66.2	46
21	40	58.8	48
22	40	50.8	50
23	41	42.7	52
24	37	34.9	54
25	31	28.1	56
26	35	21.5	58
27	24	15.6	60
28	26	10.6	62
29	21	5.9	66
30	14	2.4	70
31	3	0.7	75
32	1	0.3	78
33	1	0.1	81
34	0		85
35	0		90

TABLE 25

Shows the Number of Pupils for the Ages 7 to 17 Answering Correctly the Number of Questions Indicated in the First Column.

Questions	*Scale Score*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*
0	23	1	3	1	2	1	3	5				
1	25	2	3	3	4	1	1	0				
2	27	2	3	2	1	1	2	0	1			
3	28	3	0	6	3	1	1	0	0	2		
4	29	0	5	5	5	1	2	0	0	0		
5	29	2	5	9	6	1	2	1	2	0	1	
6	30	2	6	6	5	1	2	2	1	0	0	
7	31	0	10	6	3	5	2	2	0	0	0	
8	32	1	8	9	6	4	4	0	1	0	0	
9	32	2	10	5	5	2	2	1	0	0	0	
10	33	2	6	15	8	6	2	3	2	0	0	
11	34	2	11	20	5	4	10	1	0	1	0	
12	35	2	9	21	12	3	3	6	2	1	0	
13	36	4	14	25	12	4	8	3	1	1	0	
14	37	1	12	23	17	12	8	4	1	3	0	
15	38	2	13	21	25	15	13	12	5	2	0	
16	39	0	17	25	23	22	15	6	4	3	0	
17	41	2	17	34	24	31	18	14	4	4	0	
18	42	1	5	20	25	20	28	19	11	5	1	
19	44	3	3	20	27	32	26	26	21	3	0	
20	46	0	4	22	33	42	34	26	19	5	1	
21	48	1	4	18	25	35	40	32	28	10	2	
22	50		2	6	30	40	40	35	25	6	1	
23	52		2	6	27	32	41	42	24	9	2	
24	54		1	8	16	29	37	42	38	8	1	
25	56			3	17	22	31	46	24	16	2	
26	58			6	9	16	35	39	23	18	1	2
27	60			0	11	16	24	24	17	8	2	
28	62			2	3	13	26	25	23	5	1	
29	66				7	3	21	19	12	5	0	
30	70				2	4	14	11	7	2	1	
31	75				1	6	3	5	4	1		
32	78					0	1	1	3			
33	81					1	1	2				
Total Pupils ...		35	173	347	399	426	500	452	303	118	16	2
Total Scale Score		1190	6280	13698	17673	20332	25117	23516	15968	6125	843	116
Mean Scale Score		34	36.3	39.5	44.3	47.7	50.2	52.0	52.7	51.9	52.7	58.0
Truer Mean ...				35.	41.	46.	50.2	54.	60.			

Determination of Age Norms.—

22. A table was constructed showing the number of pupils of each age answering correctly a defined number of questions. Table 25 shows the results.

23. The total number of pupils for each age, the total scale score for each age, and the mean scale score for each age was computed. These are shown at the bottom of Table 25. The second vertical column rather than the first vertical column of Table 25 was, of course, used in making the latter determinations.

24. An effort was made to determine a truer mean scale score for each age than was yielded by step 23 above. Since few tests were given to any grade below the third it is evident that, on the whole, only the brightest seven-year-olds and eight-year-olds were tested. The less gifted were still in the kindergarten, first grade, or second grade. This factor of selection has operated to make the mean for these two ages in particular unduly high. Since no tests were given in high school except to twelve-year-olds and since the brightest pupils above the age of about thirteen have graduated from the elementary school or have left to go to work, thereby leaving the stupider children in the elementary grades, it is evident that the mean scale for each of these upper ages is too low or is unreliable because of the small number of cases.

That the factor of selection has been operating can be shown not only by logic but also by age-grade statistics. Since age-grade tables have been widely circulated by Ayres, Strayer, Thorndike, and others, repetition here is not necessary.

Not only selection but also the amount of the selection is indicated by the total number of pupils of each age shown at the bottom of Table 25. Special pains were taken to secure all the twelve-year-olds. Practically all of them had entered school. None had left school except to enter high school where they were found and tested. Five hundred twelve-year-olds were found. As shown in Table 25 the lower the

ages go below twelve the smaller the number of pupils becomes even when it is certain that there are at least as many children for each age below twelve as there are for age twelve. A similar decrease in numbers occurs above twelve.

The method of determining the truer means for ages below and above the age of twelve was to compute a median score on the rough assumption that there existed in the schools or communities 500 children of each age whether tested by me or not, and that the untested younger children were below the median of their own age group while the untested older children, particularly the compulsory-school-age children of 13 and 14, were above the median of their own age group. Since 500 minus 35 equals 465 which is far in excess of the median 250th seven-year-old child the truer mean for this age is evidently below the lowest extremity of the scale. Since 500 minus 173 equals more than 250 it is equally impossible to determine by this method the truer mean for eight-year-olds. It is possible, however, to find a truer mean for nine-year-olds, since 500 minus 347 equals 153 which is less than 250. To find the median score for this age it is necessary to count down 250 minus 153, i.e., 97. This gives a median score or truer mean of 35. In similar fashion a truer mean was computed for ages 10 and 11. The computation of a truer mean for ages 13 and 14 was simpler for here the non-tested pupils were not missing from the bottom but from the top of the group. Hence all that was needed was to count down 250 pupils. Since ages 15, 16, and 17 did not have 250 pupils a truer mean for them could not be determined.

As stated before, it is certain that the means for the lower ages are too high and for the upper ages too low. It is equally certain that the truer means for the lower ages are too low and for the upper ages too high. The method of determining the truer mean assumes that the untested younger pupils are *all* below the median of their own age group and that the opposite situation obtains for the older children. For this assumption to be absolutely correct would require,

among other conditions, that the system of promotion in the public schools be very accurate. Since we know that it is very fallible we can be morally certain that some of the young untested children would have made records above the median of their age group had all been tested. Since the number of these would undoubtedly have been comparatively few there are strong grounds for believing that the truer mean scale score is a closer approximation to the genuinely true mean for which we are searching than the mean scale score.

There are highly technical statistical procedures for inferring the location of the true mean, but inspection, guided by the mean and the truer mean, was accurate enough for the present purpose. The row of crosses in Fig. 3 pictures the means, and the row of circles the truer means of Table 25. The straight line, following the row of circles more closely than the row of crosses, gives my estimate of the position of the true mean and consequently of the true age norms.

The line for the true means shown in Fig. 3 is a straight line and has been extended above and below ages for which actual data is available. The most probable true means coincide so closely to a straight line that a straight line has been assumed in order to simplify interpolation in the computation of month standards from year standards. There is strong probability that for very young ages the line should bend downward. There should probably be a similar bend for ages beyond about 15. Proof of this last is the fact that the mean scale score for very superior teachers is just about equivalent to the standard score for age 18 if the true-mean line is continued upward as a straight line. Since, however, the line of the true means is a straight line for most of the ages it is continued as a straight line above and below until data has been secured upon which to estimate the course of the line at the two extremes.

The straight line continuation has the added practical advantage of making it possible to assign to any pupil a reading age though he makes a scale score of 90, even when

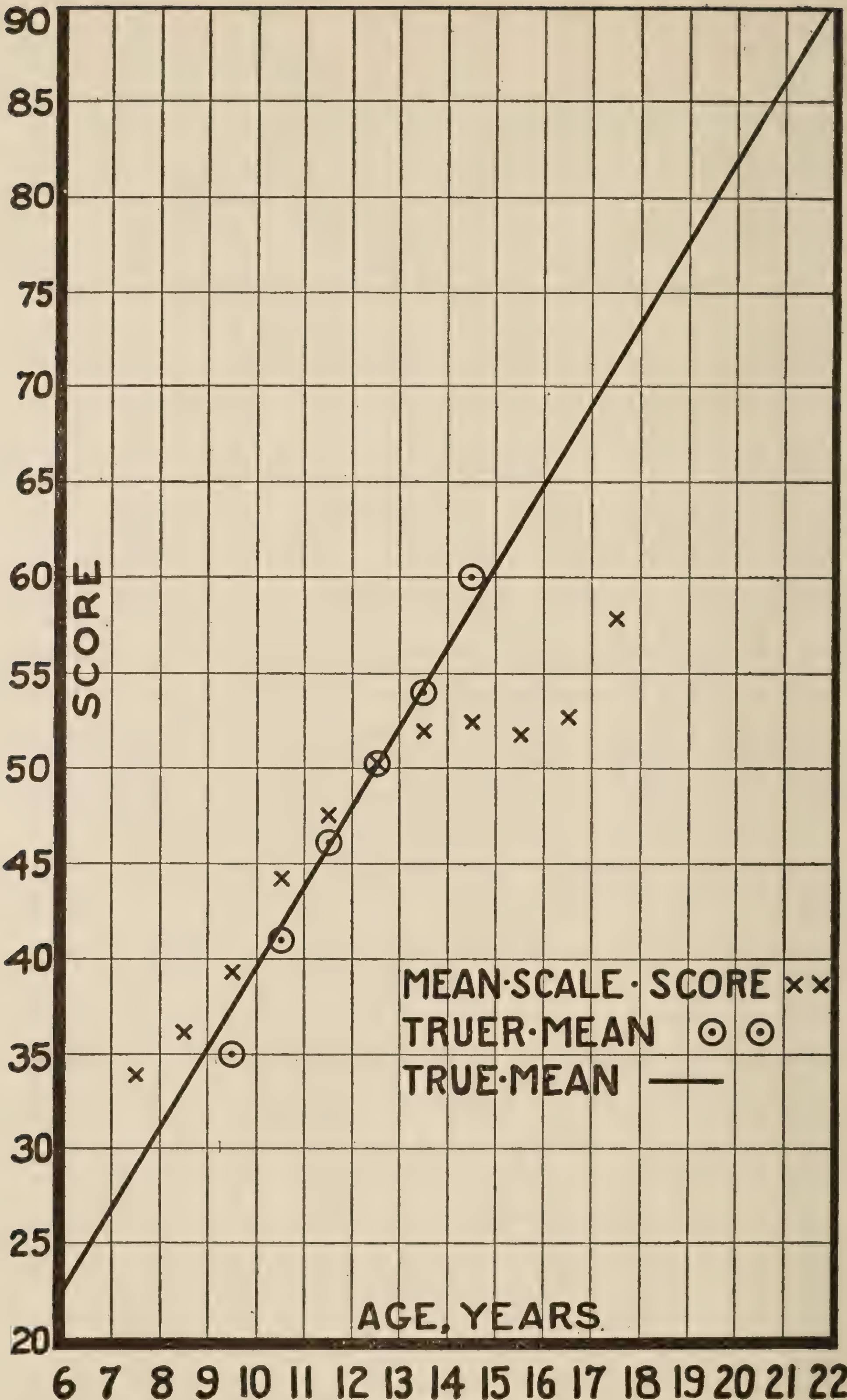


FIG. 3. Shows the Obtained Mean Scale Scores for Each Age, the Truer Mean or Estimated Median Score, and the True Mean Based upon the Obtained Mean and Truer Mean. (See Table 25.)

it is certain that no age, however old, has a true mean as high as some younger children can attain. This practical advantage does not, however, outweigh the necessity of determining empirically and reporting age standards for very young and very old children.

Norms for each month can be read from the straight line in Fig. 3. The mean and truer mean scores shown in Table 25 are, of course, for ages 7½, 8½, and so on.

Determination of Grade Norms.—

25. A table similar to Table 25 was constructed showing the number of pupils in III A, III B, IV A and so on making defined scores. A mean scale score was computed for each section of each grade. This gave norms for grade and section.

Special Extension of Scale.—Step 25 completes the scale and the establishment of age and grade norms unless it is desired to test pupils whose ability extends beyond the range of the ability of twelve-year-olds. Except in very rare tests twelve-year-old ability will go down as low as it is ever necessary to measure. In such rare traits the scale may be specially extended downward. Very few pupils will ever be found in the elementary school who possess an ability in excess of that of the ablest twelve-year-old. A few high school students cannot be given their exact scale score unless the scale is specially extended upward. Any student answering more than 33 questions correctly could not, according to Table 24, be assigned any other scale score than 81 plus. If a scale is intended for frequent use with advanced high-school students in addition to elementary school pupils it will be desirable to extend the scaling above 81.

26. Step 19 was repeated for all sixteen-year-old high-school students in a certain high-school. Students who were fourteen years old or fifteen years old might have been used instead.

27. It was determined that in this particular community 20 per cent of all sixteen-year-olds were in high school. This meant that the 200 sixteen-year-olds tested were, on the

whole, the brighter portion of the 1000 sixteen-year-olds in the community.

28. Beginning at the bottom of the frequency distribution mentioned in step 26, the number of sixteen-year-olds exceeding-plus-half-those-reaching 35 questions correct was determined. A similar determination was made for 34 questions correct, and then for 33 questions correct.

29. To get the per cent of sixteen-year-olds exceeding-plus-half-those-reaching 35 questions correct and then 34 questions correct, and then 33 questions correct, the numbers found in step 28 were divided, not by 200, but by 1000. This is an approximate correction for the failure to test all sixteen-year-olds in the community.

30. These per cents were converted into *S.D.* values by means of Table 23.

31. The scale score for 34 questions correct was 4 points above the scale score for 33 questions correct, so 4 points were added to the 81 shown in the last column of Table 30. The points difference between 35 questions and 33 questions was 9, so 9 was added to the 81 of Table 24. This gave a scale score of 85 for 34 questions correct, and a scale score of 90 for 35 questions correct, thus extending the scale upward.

A special extension of a scale downward will rarely be necessary. If so it can be done by repeating steps 26, 27, 28, 29, 30, and 31 for, say, eight-year-olds. In this case scale-score differences, according to eight-year-olds, would be subtracted from the lowest scale score, according to twelve-year-olds, instead of being added to the highest scale score as described above.

How to Increase Accuracy of Scaling.—The only substantial defect of the T scale is the tendency toward unreliability of the lower and upper scale scores. While not necessary it is advisable to correct for this defect by repeating the process shown in Table 24 for eleven-year-olds and then for thirteen-year-olds. The average of these three scale scores for each total number of questions correct will

approximate what would have been secured had three times as many twelve-year-olds been tested. If it is feared that the thirteen-year-olds are not about as much above twelve-year-olds as eleven-year-olds are below twelve-year-olds, the scale scores as of eleven-year-olds and of thirteen-year-olds may each be equated with those for twelve-year-olds before the three scale scores are combined. The procedure for equating is extremely simple. For example, suppose that the scale score for 22 questions correct is 51 for the eleven-year-olds, 50 for the twelve-year-olds, and 49 for the thirteen-year-olds. By adding 1 to the scale scores for eleven-year-olds and by subtracting 1 from the scale scores for thirteen-year-olds both series will be made comparable to the scale scores for twelve-year-olds. This is on the reasonable assumption that the variability for these three ages is approximately the same. If desired the amount to be added and subtracted can be determined more exactly by using 21, 23, 20, 24 and other number of questions correct to supplement the determination for 22 questions correct.

Publication of Scale.—

32. A leaflet was prepared for distribution with each package of test sheets purchased. This leaflet gave detailed directions as to how to apply and score tests, how to compute pupil and class scores, how to utilize age and grade norms and the like. All too frequently excellent scales are practically valueless because these essential details are given little attention or are totally overlooked.

This completes the process and the scale is ready for use. The additional technique yet to be described, namely the use of a calibrator, applies only in case two or more duplicate scales are being constructed.

Use of Calibrator in Scale Construction.—Since not one but several reading scales were made at once a calibrator, which is merely a few questions common to each preliminary and final test or group of tests, was used in constructing the reading scales. While a calibrator is never a necessity it is of value when the number of different prelim-

inary and final tests become too great to apply them all to the same pupils, or when, for some other reason, it becomes necessary to apply part of the tests to one group of pupils and part to another presumably equivalent group.

If the distribution of the ability of one group is identical with that of the other group, a calibrator is superfluous. If, however, there is a difference between the two groups, the scale difficulty of each question may differ so much that *S.D.* values are not comparable from group to group and hence questions cannot conveniently be shifted at will from one preliminary test to another in arranging the final tests. Also the scale scores for each total number of questions and the age and grade standards in terms of these scale scores will not be exactly comparable from scale to scale. The calibrator questions, since they are answered by the pupils in each group, become a means of referring all *S.D.* values, whether for each individual question or for each total number of questions correct, to one group or to the mean of the two groups.

Suppose that 49 per cent of one group of twelve-year-olds answers a certain calibrator question correctly while 52 per cent of another group answers this same question correctly. According to Table 23 the question has an *S.D.* value of 50 for the former group and an *S.D.* value of 49 for the latter group. The difference is 1 in favor of the latter group. Evidently the latter group is, on the average, composed of abler pupils. By averaging the *S.D.* differences between the groups on several calibrator questions an accurate measure may be secured of the mean difference in ability between the two groups.

Let us suppose that the mean difference turns out to be 2 in favor of the latter group. By adding 2 to the difficulty value of each question, according to the latter group, all values for the individual questions become comparable to those for the former group. Or by subtracting 2 from the difficulty values as determined by the former group these values become directly comparable to those for the latter.

Or by adding one-half of 2 to the values for the latter group and by subtracting one-half of 2 from the values of the former group the difficulty values of all questions are made what they would have been had all questions been answered by all pupils in both groups. In this way a scale constructor can get the benefit of a larger number of pupils without actually trying all of his material upon all the pupils.

The same difference that is used to equate the scale values of different questions may also be used in the same way to equate, with reference to either group or the mean of both groups, the scale scores for each total number of questions correct.[1]

The statements made in the two preceding paragraphs are only roughly true. The difficulty value of any particular question or the scale score of any total number of questions correct is a resultant partly of the mean ability of the group and partly of the form of the distribution of that ability. When one group of pupils is better as a whole or poorer as a whole than another group the calibrator will smooth out the difference. The calibrator concerns itself only with *average* differences between groups.

After the scale scores on the final scales have, in this way, been made mutually comparable, or even when no attempt has been made to make them comparable, it is advisable to average the age norms on all the scales and the grade norms on all the scales to get more accurate age and grade norms. When, however, comparability of values has been established it is really necessary to establish age and grade norms for only one of the scales.

Short Cuts in Scale Construction—Scaling Teachers' Examinations.—The process of scale construction may be shortened by the elimination of steps 3, 4, 5, 6, 7, 8, 9, 10, 11, and 12. It sometimes happens that the test questions are selected for some diagnostic reason and that, as a con-

[1] At the time of writing there is some doubt about the correctness of this statement. It is possible that a calibrator may be used as recommended to equate scale values of individual questions, but not scale scores. The problem is being studied empirically.

sequence, no questions must be eliminated because of scoring and statistical considerations. Again, it sometimes happens that the arrangement of the questions should be based upon considerations other than those of difficulty and that this arrangement can be finally decided without a preliminary try-out. In making the reading scale, for example, it was impossible to so arrange the questions that each question would be more difficult than the immediately preceding question. Each question had to be grouped with the prose or poetry selection upon which it was based. Finally, there is adequate justification for the contention that test elements need not advance with continuous increases in difficulty and that only a rough arrangement in order of difficulty is necessary.

Further, the process can be greatly simplified if the purpose is merely to construct a scale and not to determine age and grade norms. In this case all that is necessary is to call together in each school and test the twelve-year-old pupils. This reduces the process to five steps as follows:

1. Prepare the test in its ultimate form.
2. Collect and test unselected twelve-year-olds.
3. Score the test sheets and compute the total number of test elements correct for each pupil.
4. Compute the per cent of pupils exceeding plus half those reaching each total number of test elements correct.
5. Convert these per cents into scale scores by the use of Table 23 and the scale is finished.

This process is so brief and simple that teachers can use it advantageously for scaling their examinations, thus:

1. Apply the examination to the pupils in the class.
2. Score each question in the examination as either right or wrong or in any other way that may seem advisable.
3. Compute the total score for each pupil on the entire examination.
4. Compute the per cent of pupils in the class exceeding plus half those reaching each score made.

5. Convert these per cents into scale scores by means of Table 23.

6. Give each pupil his scale score instead of his original total number of points on the examination.

A teacher will find such a scale score the most convenient form in which to keep her record of the pupils. Such a score will make the pupil's record on any examination comparable with his record on any other examination. Furthermore, since the score is in statistical form it is possible for the teacher to combine records by simple addition. When records are kept in terms of A, B, C, D, E, statistical manipulation is impossible.

II. Evaluation of Methods of Scale Construction

Reference Point.—Whatever the measurement scoring must have some starting point—some reference point. Kalamazoo has a location, but the location is not very intelligible to anyone unfamiliar with Kalamazoo unless given some reference point or points. If we say Kalamazoo is so many degrees west longitude and so many degrees north latitude, the reference points are the line of longitude passing through Greenwich and the line of latitude corresponding to the equator. According to scientific measurement the reference point for measuring an individual's height is either the soles of the bare feet or the actual crown of the head. Whether the thing measured be distance, time, weight, courage, reading ability, or arithmetical skill, there must be a starting point for scoring.

The following drama will illustrate the need for a commonly understood reference point:

TRAGI-COMEDY OF ERRORS

ACT FIRST

Railroad Station, Richmond, Va.

Enter Traveler, Native of Baltimore, Native of Savannah, Bostonian and Author of "HOW TO MEASURE IN EDUCATION."

Traveler: Is New York City farther than Philadelphia?

Author: Define your point of reference. (*Exit* Author.)

Native of Baltimore: Yes.

NATIVE OF SAVANNAH: Yes.
BOSTONIAN: No! (*Exit* BOSTONIAN.)
TRAVELER: How much farther is New York City than Philadelphia?
NATIVE OF BALTIMORE: About twice.
NATIVE OF SAVANNAH: About one-tenth!

The End

Scientists soon discovered that scientific progress was handicapped by the fact that different individuals were using different reference points when measuring temperature. Finally after long wasteful delays two competing reference points have been adopted, one which places the zero point of the temperature scale 32 degrees below the freezing point of water and one which locates it at the freezing point of water. In similar manner scientists agreed to make the zero for the height of land forms the sea level. They could have made it the center of the earth or the base of the Acropolis. In the measurement of many things in life then there is no one point divinely called to be zero. Convenient zero points have been proposed, debated, and arbitrarily adopted.

Mental measures have for years been searching for an appropriate reference point or points. The tendency has been to search for some absolute zero point for the trait being measured. This has resulted in a different zero point for each scale made. If the process continues we shall have hundreds of zero points each of which is extremely nebulous, and no one of which is generally accepted. The resulting confusion would enormously handicap the development of mental measurement.

We have had not only a different reference point for each test, but different methods of locating this point. *First:* the reference point on unscaled tests is just no score on the material of the particular test. *Second,* the reference point on certain scales is a zero point guessed at by the author of the scale. *Third,* the reference point on other scales, particularly judgment scales, is the median judgment of judges as to the location of zero merit in composition. *Fourth,* the reference point on other scales is a zero point located by the

use of the per cent of pupils in some early grade who make no score on very easy material. *Fifth,* the reference point for other scales is 3 *S.D.* below the mean of the group for whom the test was devised. *Sixth,* the reference point on another scale is simply the lowest score made. Still other methods of locating reference points have been used.

Since few mental measurers agree as to the best method of locating a zero point, since few agree as to just what the zero for reading or any other mental trait is, since any such point if actually found is bound to be relatively invisible and hence more or less valueless as an aid to the proper interpretation of scores, since prevailing methods of locating zero is certain to produce as many different points as there are scales, and since this last must inevitably result in general confusion, this book proposes that a common reference point be arbitrarily adopted for all tests which are to be used in the elementary and perhaps high school. It is recommended that this reference point be not a zero point at all but instead the mean performance of children between the ages of 12.0 and 13.0. Such a reference point could be used for any mental trait regardless of the location of its absolute zero, if such there be.

The method of scale construction described in the preceding section makes the mean performance of twelve-year-olds the point of reference. Any pupil who makes a scale score of 50 has an ability equal to the mean ability of twelve-year-olds. According to Table 23 his ability is exceeded by 50 per cent of twelve-year-olds. Any pupil who makes a scale score of 40 has an ability which is 10 units or 1 *S.D.* below the mean ability of twelve-year-olds. According to Table 23 he is exceeded by 84.13 per cent of twelve-year-olds. Any pupil whose scale score is 75 is 2.5 *S.D.* above the mean ability of twelve-year-olds. Only a little more than six-tenths of one per cent of twelve-year-olds exceed this record.

Actually in a mathematical sense the zero of the scale is 5 *S.D.* below the mean of twelve-year-olds. This is a case where the zero of the scale is not the actual reference point.

The mathematical zero was located at 5 *S.D.* below the mean instead of at the mean for three reasons. *First,* this procedure eliminated cumbersome plus and minus signs. Had the zero been placed at the mean it would have been necessary to report a pupil's score as — 3 *S.D.*, or + 2.5 *S.D.*, or the like. It is much more convenient to report a pupil as 20 or 75, or the like. *Second,* this procedure gives a convenient range of points between 1 and 100 and locates the reference point at the easily remembered 50. *Third,* this procedure carries the scale down as far and up as far as anyone will need to go. This would not be the case if the scale ranged only from — 3 *S.D.* to + 3 *S.D.* *Fourth,* this procedure gives a mathematical zero which is close to the supposed absolute zeros for reading, writing, spelling, composition, completion and other typical mental functions. An investigation of several scales which have been constructed by Buckingham, Trabue, Woody and others showed that the zero points selected by them were only slightly above 5 *S.D.* in terms of twelve-year-olds below the mean of twelve-year-olds, thereby compensating for the general feeling on the part of most of these authors of scales that their zero points were probably a bit high. In sum, 5 *S.D.* below the mean of twelve-year-olds is itself a reasonably good zero point for many commonly measured abilities and in addition serves to accentuate an even better reference point as well as serving other practical purposes.

Which sort of reference point should be used depends in part upon the use to be made of the measurements. It is generally acknowledged that the best way to appreciate the ability of a pupil is to refer him to the mean ability of his own or some standard group. On the other hand, it is acknowledged that the best reference point to employ when the *times* statement is made is the absolute zero of the trait in question. The absolute zero is desirable, for example, when it is desired to state that James has two times the ability of John. The proposed actual reference point meets almost perfectly the needs of the former group. The proposed

mathematical zero at 5 *S.D.* below the mean of twelve-year-olds also meets the needs of the latter group reasonably well, for 5 *S.D.* below the actual reference point is an approximate absolute zero for most school traits. A slight concession on the part of those who prefer an absolute zero will make possible a uniformity which, in my judgment, more than compensates for the loss entailed in making this conccssion.

Another objection which may be urged against the mean of twelve-year-olds as a reference point is that any score above or below this point would not indicate whether the pupil possessed much or little of the trait being measured, for the mean twelve-year-old pupil possesses relatively little of certain traits and relatively much of certain other traits. The only cure for this defect is to use as the reference point the undiscoverable absolute zero of the trait and, in addition, to lose one of the values of the proposed reference point. If the absolute zero point is found and used no one score can always indicate the mean ability of twelve-year-olds. It is probable that is is easier for teachers to estimate the amount of a trait possessed by a pupil from a knowledge of how much a mean twelve-year-old pupil possesses than from a knowledge of a score's distance above some theoretical absolute zero. Furthermore, the material of the test itself is the best common sense index of the amount of ability possessed by anyone taking it.

The only other convenient reference point which has been proposed is the time of birth. Such a reference point is used in the Stanford Revision of the Binet-Simon Scale. In fact, its use has been rather general in the case of intelligence scales which are true scales. The chief objection to the universal adoption of this popular and easily understood reference point is that it practically compels the adoption of a measuring unit which, as will be shown later, is less satisfactory.

Unit of Measurement.—Just as all measurement requires some reference point so all measurement requires

some unit. The reference point for a mountain is sea level. Its height above this reference point is expressed in terms of a certain measuring unit called a *foot*. The reference point for measuring time is the birth of Christ, or January 1st, or 12 M., and the units are centuries, years, days, hours, minutes, and seconds.

The variety of reference points is almost equaled by the variety of units for mental measurement. The reader can get some idea of the situation by remembering his own "buzzing blooming confusion" upon being suddenly asked to learn the tables for meters, liters, grams, pounds, marks, shillings, yards, dollars, etc. This confusion multiplied tenfold is a sample of the present chaos in mental measurement. Some employ simple scoring units, others employ as a unit some function of the variability of any one grade, others employ some function of the variability of all the grades treated separately and then combined through the determination of inter-grade intervals, others employ some function of the variability of all grades combined, others employ the variability of judgment, others employ the variability of adult performance, others employ relative worth, and others employ still different units. This book proposes that a single common unit of measurement be adopted for all mental scales to be used in the elementary school, namely, some function, preferably *S.D.*, of the variability of twelve-year-old pupils. It is further proposed that a similar unit based upon, say, sixteen-year-old pupils, be used for tests designed especially for high schools. Tests designed especially for the primary grades could be scaled for, say, eight-year-olds.

The highest point in the technique of mental scale construction has been reached by Thorndike and his students. They have used some function of variability as a unit and this is admirable. They have used the variability of a grade, which is not so admirable. Many forces are at work such as reorganizations of grade systems, improvements in classification, and the like, which are bound to profoundly alter, in a relatively short time, all scale values and the sig-

nificance of the scale units employed. Any unit based upon such an artificial and ephemeral group as a grade lacks the necessary permanence. They have used values based upon the variability of several grades and have combined these values through an elaborate procedure of weighting values and determining inter-grade intervals. This procedure has the merit of giving temporarily reliable results, but the whole procedure is altogether too laborious for it to be generally used. Furthermore, the values when pooled for several grades with intricate weightings cease to be interpretable. The only sort of variability which has much meaning is the variability of some one defined group. Even so this group of scientific workers has done more to further the cause of accurate scale construction than any other group in the world.

The other high point in scale construction began with Binet and Simon and culminated in the Stanford Revision of the Binet-Simon Scale by Terman. This line of development has been popular rather than technical. Its reference point has been the time of birth, and its unit of measurement has been one year of growth or some subdivision thereof. These are reference points and scoring units which all can understand. They utilize chronological age, one of the most abiding features of human life.

There are, however, some very serious objections to this unit of measurement. While a permanent one, it is not equal in the truest sense of the word, at all points on the scale. A fact, now taken for granted, is that the interval between 8 and 9 years of age is larger than the interval between 14 and 15, in the case of intelligence and probably for many school traits as well. Furthermore, in the case of certain mental traits, the units become of zero size beyond about sixteen years of age. In abilities where a loss occurs, after instruction in the elementary school ceases, the age unit may be actually less than zero, i. e., negative. Finally, because of the late entrance into school of some pupils and because of the disappearance into the social medium of a

goodly per cent of the graduates and over-compulsory school-age pupils of the elementary school, it becomes difficult, if not impossible, to build up a scale below an age of 8 years and above an age of 12 years. This means that such a restricted scale cannot satisfactorily score a very poor or very able pupil. To accurately extend the scale so it will measure these individuals requires that a test be previously scaled beyond these points by some other method of scaling. Lending itself as it does to easy interpretation and to the ready computation of quotients, the age unit is deservedly popular. It is, however, most serviceable in the realm of intelligence measurements where, presumably, retrogressions do not normally occur until senility sets in. Its defects are almost fatal for universal mental measurement. Even so it is at least the second best unit for universal adoption.

The unit proposed for universal adoption by this book, namely one-tenth *S.D.* of twelve-year-old children has long been used by careful mental measurers to compare pupils with other pupils in their own age group. This unit is equal at all points on the difficulty scale, which is the chief characteristic of the unit employed by Thorndike and his students. It is based upon chronological age which is the chief characteristic of the work of Terman and his predecessors. It is a function of the variability of a defined group and a group which is easily located. A scale which uses this unit reaches as low and as high as the ordinary requirements of practical measurement. Special extension at the top or bottom is a simple process. And not of least importance is the fact that the construction of the scale which employs this unit is not particularly laborious. The subsequent improvement of scale values is simpler still. In sum the proposed unit combines most of the virtues and eliminates most of the defects of the two chief contemporary methods of constructing mental scales. In a certain sense it unites the two great lines of scale development. Since the two greatest contemporary exponents of these merging methods

are Thorndike for the one and Terman for the other, it is a tribute to their genius to call the proposed unit, namely one-tenth *S.D.* of unselected twelve-year-old children, a Thorndike-Terman, or, for brevity, a T.

This variability-of-an-age-group unit has long been used among psychologists. But they have used the unit for purposes of comparing a pupil with other pupils in his own age group. Table 23 will facilitate such use. The greatest need, however, is to compare, in some quantitative fashion, a pupil in one age group with another pupil in another age group or the mean of one age group with the mean of another age group. This need is met by employing the T for twelve-year-olds as has been done in this chapter.

The proposed unit and method of scale construction is applicable in almost if not all situations where mental traits are being measured. As has been seen no obstacle has appeared when the object was to make a scale to measure how much difficulty a pupil could achieve. Instead of difficulty the basis could just as well have been relative worth or frequency of occurrence or something else. Any sort of a basis which somehow yields varying scores for pupils may be used.

Even product scales may be transmuted into T scales, thereby making all scales performance scales. As has been pointed out product scales such as those of composition and handwriting are really nothing more than scoring instruments. Answers to reading questions may be scored as either right or wrong, but a whole reading test could not conveniently be scored right or wrong or even 2, 1, or 0 all at once. A composition test is scored all over all at once. Those doing the scoring required some scaled scoring instrument. Product scales are such scoring instruments. After the scoring an additional step is possible and may be desirable, namely, to take the scores of twelve-year-olds based upon product scales and scale them as shown in Table 24. The T scale should be used for all performance tests. The method used by Hillegas should be used in making all

product scales. In order that comparisons between performance on all tests may be easily made, product-scale scores should be converted into T-scale scores.

Method of Combining Scoring Units.—Most tests and scales, excepting such scales as Monroe's Standardized Reasoning Test in Arithmetic and Henmon's [2] French tests, compute a pupil's score by getting a simple total of the units satisfactorily done. A pupil's score on Courtis' test in addition is the total number of examples covered. His score on Woody's test in addition is likewise the total number of examples done satisfactorily, which, at the same time, indicates the total number of *P.E.* difficulties surmounted. His score on Thorndike's test of visual vocabulary is a measure of how far up a scale of difficulty he can work, i. e., his score is a simple total of the *P.E.* distances covered. His score on the Nassau Composition Scale or Ayres' Handwriting Scale is a simple total of the units of merit his own merit exceeds.

Among these scales which combine units by simple total there are at least three distinct ways of doing it. There is, first, the product-scale method. The scorer slides the pupil's specimen of handwriting, say, up the scale until its proper location on the scale is determined. The pupil *certainly has an ability beyond that of any lower position on the scale.*

But how is it with, say, the Woody test in addition or the Trabue Language Completion Scales? Does an ability to do an example with a *P.E.* difficulty of 4 guarantee an ability to do any other example with a *P.E.* difficulty less than 4? Not at all. It frequently happens that a pupil succeeds with an example whose difficulty is greater than that of examples on which he failed. The case might occur of one pupil who does the three easiest examples and misses all the rest, and of another pupil who does the three most difficult examples and misses all the easier ones. According to this second method both pupils would make the same score. This

[2] V. A. C. Henmon, "Standardized Vocabulary and Sentence Tests in French"; *Journal of Educational Research,* February, 1921.

sort of thing frequently happens in a less drastic fashion. However infrequent such events, these gaps are not exactly chasms of music to students of the theory of measurement.

There is, third, the 20%-error method of Thorndike's and Haggerty's reading scales, and the 50%-error method of Woody's Fundamentals of Arithmetic Scales, Series A. These scales locate a pupil's or a class' score at that point on the scale where 20% of error is made or 50% of error is made, and to a certain extent wink at what has happened below this point and above this point.

Of these three, the method of the product scale is the only thoroughly satisfactory one. Whether we attribute it to the accidents of chance or to the "iron laws" of nature, a performance scale cannot be so constructed that pupils will conveniently work every test element in order of its difficulty and then suddenly break off. The difficulties of test elements, as determined by testing a large number of pupils, may not exactly fit any individual child.

Of the last two methods mentioned, the one used by Woody and Trabue, namely, the number of examples done correctly, or the number of sentences correctly completed, is a far more convenient method than the 20%-error method. Why, then, was the latter used at all? In the first place, the method used by Woody and Trabue requires a scale made up of equal steps. Woody used the 50%-error method for his Series A scales because they do not progress by even steps. The 20%-error or 50%-error method is of such a nature that a scale of unequal steps does no harm. In Table 26 are two scales, the first with, and the second without, equal steps. The "Am't to be added" tells how much a pupil's score should be increased when he does the test element correctly.

If, in the first scale, a pupil misses test element C he fails to get the one point increase, if he gets C correct his score is increased by one point. To lose or gain anywhere means a loss or gain of just one point. In the second scale a loss or a gain would be one point if he missed or worked A, B, C

TABLE 26

A Comparison of Two Scales, One of Which Progresses by Equal Steps and One Which Progresses by Unequal Steps

First Scale Elements	A	B	C	D	E	F	G	Maximum Score
P.E. value above zero...	1	2	3	4	5	6	7	
Am't to be added.......	1	1	1	1	1	1	1	7
Second Scale								
P.E. value above zero...	1	2	3	3.1	4.5	5	6	
Am't to be added.......	1	1	1	0.1	1.4	0.5	1	6

or G or a loss or a gain of one-tenth or one-and-four-tenths or five-tenths if he missed or worked respectively D, E, or F.

To some it does not seem reasonable to penalize a pupil 1.4 for missing E and only 0.1 for missing D. In other words test element E has fourteen times as much influence upon a pupil's score as test element D, although the casual observer would not remark any special superiority in element E. Again, when the steps of the scale are very irregular, the resulting scores may cause an injustice in comparing one pupil with another. Here are the scale increments and total score earned by two pupils:

									Score
Pupil *a*	1	1	0.1	0.2	0.1	0.3	0.2	0.0	2.9
Pupil *b*	1	0	0.0	0.0	0.0	0.0	0.0	0.0	1.0

Pupil *a* appears to be nearly three times as able as pupil *b*. As a matter of fact their ability may easily be closer together than the scores indicate. Pupil *b* could not quite do the second test element. Had there been several test elements with difficulties between 1 and 2 just as there is between 2 and 3, pupil *b* might have made a score of 1.9. In sum because of the irregular steps pupil *a* was specially favored. The favoritism is, of course, not serious, but it definitely exists. The percentage-of-error method, especially an elaboration of it used by Kelley, takes care of this irregularity.

Monroe's Standardized Reasoning Test in Arithmetic may be selected as a sample of those tests which combine scoring units according to the method of cumulative total. Table

27 shows the difference between the methods of simple total and cumulative total.

TABLE 27

A Comparison of the Method of Simple Total and Cumulative Total for Combining Scoring Units

Test Element	A	B	C	D	E	F	Maximum Score
P. E. value above zero............	1	2	3	4	5	6	
Am't to be added (simple total)....	1	1	1	1	1	1	6
Am't to be added (cumulative total)	1	2	3	4	5	6	21

The difference between these two combining methods can be shown again by the following analogy—the computation of a man's height. As the illustration shows, the cumulative total is a rather Brobdingnagian one.

	Sole	Ankle	Knee	Hip	Shoulder	Crown	Total Height
Scale value.....	0	3 in.	22 in.	41 in.	61 in.	74 in.	
Simple total...	0	3	19	19	20	13	74 in.
Cumulative total	0	3	22	41	61	74	201 in.

Another illustration does not present the cumulative total in such an unfavorable light. In this illustration a boy has agreed to carry some wood and is to be paid by the number of pounds of wood he carries. Here the cumulative total appears the natural one to use.

	Stick I	Stick II	Stick III	Stick IV	Total
Scale value	10 lbs.	15 lbs.	20 lbs.	35 lbs.	
Simple total	10	5	5	15	35
Cumulative total.	10	15	20	35	80

Thus the two methods measure different things. The simple total measures the height of the ability while the cumulative total measures the number of work units completed. The simple total measures how heavy a log a child can carry. The cumulative total measures the weight of the logs he does carry.

Each method is best for its purpose, but the cumulative total does not appear to be specially advantageous in educational measurement. Practically the world thinks of ability

as yielded by the simple total. In comparing the ability of two persons the customary mental act is similar to a comparison of their heights. Properly to interpret differences between scores computed by the cumulative method, educators must change their customary mental set. They must remember that if one pupil has a score of 4 and another has a score of 10, that the second pupil has done two and one-half more work units than the first pupil. It does not mean that the second pupil is two and one-half times as able as the first one. If this point of interpretation is kept in mind it makes little difference which method is used, for while scores computed by the two methods are not proportional they do correlate closely, especially if the test in each instance measures the maximum ability of the pupils.

The use of the simple or the cumulative total for measuring how difficult material a pupil can work should not be confused with the problem of measuring speed. In most speed tests every test element requires approximately equal time for the pupil to complete it. Hence the units are in that sense equal and a pupil's speed is determined by the number of test elements completed or tried per unit of time. But occasionally the convenient work units in a speed test require quite varying times. In such instances it is necessary to determine the relative amount of time required for each work unit, i. e., to give each work unit a time value. It is quite legitimate to add up the rate values of all the test elements completed by the pupil in order to get a measure of his speed of work. But it should be observed that the rate value on each element does not reach back to the time of the starting of the entire test.

In sum, we have seen that, with the exception of product scales, no scales have a totally satisfactory method of combining scoring units to make a pupil's score. The 80–20 per cent method devised by Thorndike is accurate enough for the computation of class scores but its use for this purpose requires a laborious tabulation by test elements. It is unsuited to the computation of individual pupil scores. An

elaboration of the 80–20 per cent method by Kelley is sufficiently accurate for the computation of pupil scores but it is too elaborate for use except for research purposes. Few non-technically trained individuals can understand or use the method. To overcome these handicaps, Trabue and Woody constructed scales the adjoining elements of which were equally far apart in difficulty. This permitted the computation of pupil scores by the simple process of addition. But to secure such a scale meant that most of the original test material had to be thrown away as useless. Monroe employed an easy method of combining units which did not require equal steps of difficulty. But as we have seen the method of the simple total appears preferable to his method of the cumulative total.

Thus the T-scale method was developed not only to provide a more satisfactory reference point and unit of measurement, but also to provide a method of combining scoring units which yields a genuine scale score for each pupil, which combines units by the method of simple total, which preserves all the original test material, and which is simple enough to be used by non-statistically trained educators. All these objects were attained at one stroke by scaling the total score.[3] Previous methods of scale construction have been to scale the test elements and then provide an additional technique for computing a pupil's score from his success in dealing with these test elements. The proposed method makes this second step unnecessary. Scaling the total number of questions correct or, when more than one point is given for each question, the total number of points made shows immediately the scale score corresponding to each total number of points, which in turn is secured by merely adding the points made on the different test elements.

Combining scoring units by scaling the total score has still another advantage. The preparation of duplicate forms of a test is an extremely tedious process when conventional

[3] I have just learned that Pintner is, by a somewhat different process, scaling the total score for his new tests. He is constructing a separate scale for each age rather than a scale which cuts across the different ages as the T scale does.

methods of scale construction are employed. For scales which secure a pupil's score by totalling the number of test elements correctly done, it is necessary to have each test element in one scale exactly matched by a test element in the duplicate scale of equivalent difficulty. Such an exact equivalence is not required in T scales. One scale may serve as a duplicate for another even where there is a considerable difference in difficulty between them.

CHAPTER XI

DETERMINATION OF RELIABILITY, OBJECTIVITY, AND NORMS

I. Reliability

Sources of Unreliability.—By reliability of a test is meant the amount of agreement between results secured from two or more applications of a test to the same pupils by the same examiner. Perfect reliability obtains when an identical examiner applies two identical or exactly duplicate tests according to an identical procedure to identical pupils. This last sentence indicates in brief those attributes which are essential to high reliability in a test, and the absence of which makes for unreliability.

The first source of unreliability in a test is variation in the behavior of the examiner produced by causes external to the test itself. There are a host of causes which have the power to produce large or subtle changes in the personality and behavior of the examiner, which behavior may in turn operate to raise or lower the pupil's scores. Such possible causes are an obstreperous pupil, a welcoming smile from the teacher, an indigestible lunch, etc. Chance might produce an especially favorable combination of causes at the first testing and an unfavorable combination at the second. Such a situation would tend toward differences in results and hence toward unreliability. This cause of unreliability is not an attribute of the test itself.

A second source of unreliability in a test is variation in the behavior of the examiner produced by causes inherent in the test itself. These causes may be in the instructions for the test, the method of scoring or the statistical treatment of

results. Perhaps the most important of these causes is inadequate description. Ideally the author's description should reveal exactly how the examiner is to deal with every significant situation which may arise in the process of testing, scoring, etc. When an author begins the description of how to administer his test, in this fashion: "See to it that all pupils understand what is expected of them," there is offered an opportunity for wide variation between different administrations of the test. *Instructions are a part of the test* and should be just about as definite and uniform as the test material itself. Definite instructions to the examiner as to how to score with uniform rigor and how statistically to treat results are no less important. A study of the extent to which Binet Test examiners have found it necessary to carry standardization of procedure will give a good idea of the importance of this point.

A third source of unreliability in a test is the never-ceasing moment-to-moment variation in pupils themselves. Like the examiner, each pupil is at any one moment influenced by a multitude of minute forces which pulse and play like mirrored lights on moving water. An automobile horn, the lonesome howl of Jack's dog, the bleating of Mary's lamb, a sudden thought of the swimming hole, growing discomfort of a strained posture, these and a thousand other large and small internal and external influences register themselves in the pupils' scores. It is rare for the registration to be equal at two test periods, and as a consequence, results from two tests differ. It is this difference which makes the test unreliable, for there is often no reason to believe that the pupils' reactions at one test period are more typical than at another.

It is not, however, always fair to judge a test's reliability by the absolute similarity between the two scores for each pupil. There are certain *constant* causes which operate to produce absolute differences in results and hence make a test's reliability appear less than its real reliability. These constant forces must be eliminated or allowed for before the real reliability can be determined. Such constant causes

are improvement due to experience with the first test and due to normal growth in the measured trait. For pupils insist upon changing with increased age and increased experience. Every second leaves its ever so little deposit. Goaded by this distracting refusal of pupils to remain stationary, Ayres has suggested that chloroforming experimental pupils would be a great convenience!

How may these constant causes be eliminated? Four methods have been employed. The methods of optimum interval, duplicate test, experimental allowance, and self-correlation. The first three methods aim to reveal the absolute similarity between the two scores for each pupil; the last method only permits a relative comparison. The optimum interval method is to allow just that interval between the first and the second test which brings the ability of the pupils most nearly to their ability at the time of the first test. A zero interval is impossible except for determining the reliability of supervisory observations and the like. A familiar law of nature forbids the application of two tests to the same pupils at the same time. Besides it is desirable that the two tests be given at different times to discover whether any pupil's score is influenced by a temporary headache or other cause of an "off day." The longer the interval the more any practice effect disappears. The interval must not be too long or the decrease in the trait due to forgetting will be counterbalanced by an increase due to greater maturity.

In choosing the optimum interval many factors should be taken into consideration. The increase due to maturity takes place less rapidly for most traits than the decrease due to forgetting. Again, some tests are of such a nature that one pupil cannot communicate to another the ability to do the test successfully, nor does any pupil retain after a brief interval any effective memory of how to do the test. By the proper juggling of these factors a pupil's ability in many tests may be practically returned to his original ability.

The duplicate test method aids the optimum interval method. The use of a duplicate test the second time partially avoids in the case of rate tests, and almost completely avoids, in the case of the difficulty tests, any increase in score due to practice effect.

The experimental-allowance method is to determine experimentally, by using a comparable group, and allow for the influence of all these constant causes for a given interval of time.

The fourth and most convenient method of all for eliminating these constant errors is that of self-correlation. It may be used alone or may be aided by both the optimum interval and duplicate test methods. The method of self-correlation is to compute the coefficient of correlation between the two series of scores secured from two administrations of the same or duplicate tests to the same pupils. If this correlation is zero, the test has no reliability whatsoever and the test is worthless no matter how many other good qualities it may possess. The nearer the coefficient approaches unity the nearer the test approaches perfect reliability. A later chapter shows how to compute this self-correlation. The change in pupil scores due to practice effect or normal growth does not deflect correlation from unity toward zero provided the influence of these factors is equal for each pupil, which is substantially the case after any reasonable interval.

Methods of Increasing Reliability.—How high should the reliability of a test be? The answer is: the higher the better. If the self-correlation coefficient is zero, the test is worthless; if the coefficient is unity, the test reliability is perfect. Here are the reliability coefficients for five standard educational tests: .55, .7, .75, .8 and .9. All uses of test results are based upon pupil scores, and a class score which is usually a mean or median of the pupil scores. A class score for a class of ordinary size will be sufficiently reliable for most purposes even though the test's self-correlation is as low as .55. But if the test scores are to be used

to make important judgments concerning individual pupils, the self-correlation should be above .9. Scores for individual pupils have some value even when the self-correlation falls considerably below .9. A test whose self-correlation is anywhere above zero is better than nothing at all for measuring individual pupils.

How may a test's reliability be increased if it falls below what is required for the purposes of the investigation? Suggestions have already been made as to how to decrease variation in the examiner and hence increase reliability through a better standardization of test procedure. An additional source of unreliability is the variation in pupils due to the operation of chance causes other than those contributed by the examiner.

There are three ways in which these chance causes of unreliability may be overcome: first, by increasing the length of the test; second, by averaging results from repetitions of the test or the test and its duplicates, and third, by a combination of the first and second methods. Unfortunately there is a limit to the number of times an identical test may be repeated owing to its increasing familiarity to the pupil, and this limit varies for different kinds of tests. In case a high reliability is desired, the existence of duplicate tests may therefore become an important factor in determining a test's worth. Duplicate tests are equally useful in preventing coaching. Just how many times a test or duplicate tests must be administered to secure a desired reliability, or just what would be the reliability of any given number of applications of duplicate tests may be quickly determined by the use of Spearman's self-correlation formula, whose solution is explained in a later chapter.

II. Objectivity

Importance of Objectivity.—A test is perfectly reliable when identical results are secured from two applications of a test to the same pupils by the *same* examiner.

A test is perfectly objective when identical results are secured from two applications of the same test to the same pupils by *different* examiners. A test is perfectly subjective when no two examiners agree. Ordinarily the objectivity of a test is lower than its reliability due to the addition of a new cause of variation, namely, the difference in the personal equation of the different examiners. Some tests are more objective than others. A test of an individual's temperature, pulse, blood-pressure, finger-length, head-circumference and the like, is usually much more objective than a test of his handsomeness or charm. Estimation of a man's height is rather subjective. The use of measuring instruments here as well as in education tends to increase objectivity. Tests are not totally subjective or totally objective. Objectivity, like reliability, is a matter of degree. Tests occupy points on a subjective-objective continuum with perhaps none located at either extreme. The degree of agreement by different examiners is the measure of a test's location on this subjective-objective scale.

Objectivity is an extremely important consideration in the construction of tests. So important is it that there is little exaggeration in stating that this criterion of objectivity is the mother of scientific educational measurement. For educational tests are an outgrowth of the extreme dissatisfaction with the subjectivity of previous methods of measuring the educational output. Progress in all sciences has been attended by a decrease in the personal equation through improvements in measuring instruments. *Verification* is the greatest word in the language of science. Education has been and still is to a large extent saturated with the personal equation. All progress in the development of education as a science is closely linked up with the creation of measuring instruments or measuring methods whose application yields verifiable results.

How to Determine and Increase Objectivity.—A test's objectivity may be determined by the solution of the following formula:

Objectivity = reliability — personal equation.

Objectivity is, however, more easily computed directly. A direct determination is made in the same way reliability is determined, namely, by comparing the absolute similarity of the two examiners' results, or by computing the coefficient of correlation between the two resulting series of pupil scores. The same precautions against the same constant errors must be taken in determining objectivity as in determining reliability. Hence the method of self-correlation will prove as convenient for computing objectivity as reliability.

How may a test's objectivity be increased? The problem in education is no whit different from the problem in other sciences. The first step in its solution is to do everything possible toward increasing the reliability of the test according to the methods sketched in the previous section. The second step is to determine, wherever possible, the amount and direction of the personal equation of the different examiners, and to allow for them. For some time to come improvement in reliability will be the most convenient and promising method of improving objectivity.

As with reliability, objectivity can be increased by a careful standardization of the entire testing process. If two examiners apply the test in different ways, disagreement is assured. If the method of scoring leaves room for the exercise of much judgment, disagreement is almost certain to arise. If there is a variation in the statistical method of computing pupil or class scores, it is hardly reasonable to expect results to agree. Adequate description can avoid most of the variation due to test application and statistical treatment. Much ingenuity is now being applied to developing completely objective means of scoring pupils.

III. Norms

Factors Influencing the Worth of Norms.— There are in use two kinds of norms or standards which need to be

distinguished, namely, standards *of* achievement and standards *for* achievement. The former means actual average achievements of age or grade groups, whereas the latter refers to goals or objectives for these ages or grades. At the last meeting of the National Association of Directors of Educational Research it was recommended that the former be called *norms* and the latter be called *standards.* When the objectives have been accepted by some authoritative national organization they may be called *standards.* If the objectives represent the opinion of the author of the test, they should be designated as *author's standards.*

Norms are more valuable when they are representative of the group with whom it is most desirable to make comparisons. If but one norm could be had, all would agree that this should be the norm for all pupils in the country. To secure this norm does not require us to test every child in the country, but it does require us to select the pupils to be tested in such a way that our sampling will show the correct proportion of each level of ability. There would be more pupils tested in schools with an average social environment than in schools with extremely poor and extremely good environments, because the average type of school is more numerous than either of the others. Simpson,[1] who desired nothing more than a satisfactory average standard for adults, tested some unusually intelligent individuals and some equally unintelligent individuals. The average of the scores made by these two groups will give a fairly accurate central tendency for the total group of which these two groups represent the extremes. If we wish norms for the first percentile, second percentile, tenth percentile, twentieth percentile and so on to the 100th percentile, Simpson's method is inadequate. We need to test pupils selected so as to represent the proper proportion of each ability level. If ten per cent of the total pupil population are at a certain ability level, then ten per cent of the

[1] B. R. Simpson, *Correlations of Some Mental Abilities;* Bureau of Publication, Teachers College, Columbia University, N. Y. C., 1912.

pupils tested should be at this same ability level, and similarly for other levels.

Norms are more valuable when they are stable. The stability of a norm is a function of the satisfactoriness of the sampling and the number of cases. What constitutes a satisfactory sampling was discussed in the preceding paragraph. Given this, or as nearly this as is humanly possible, the greater the number of cases the greater the stability of the norm. This statement is not exactly true, for as Pintner and Paterson [2] point out the perpetual multiplication of cases is not necessary to secure a stable norm. There comes a point where the addition of new cases does not materially influence the previous determination. What we really want answered is: how many cases are necessary to establish a reasonably stable norm? Until we have collected more experience, there is, as Pintner and Paterson state, no way to tell except by averaging the scores of varying numbers of cases and by watching the resulting fluctuations in the averages. When the addition of, say, 100 new cases does not materially alter the previously determined norm, the norm has stabilized.

The number of cases required to stabilize a norm varies with the type of norm being stabilized. There are in common use three kinds of norms: (a) the average performance norm, (b) the placement-of-a-test-at-a-specific-age-on-an-average-scale norm, and (c) percentile or variability norms. The placement of a test element in a grade scale is akin to the establishment of percentile norms. While there is a slight variation in practice, the usual method of placing a test on such an age scale as the Binet-Simon Intelligence Scale and the Pintner-Paterson Performance Scale is to place a test at that age level where 25% fail it and 75% get it correct. These per cents were selected on the principle that the 25% poorest should fail the test and the 50% normal and 25% supernormal should pass it. Now since more of

[2] Rudolf Pintner and Donald Paterson, *A Scale of Performance Tests;* D. Appleton & Co., N. Y., 1917.

the pupils taken in the sampling will fall near the median than near the 75 percentile, and since more will fall at the 75 percentile than at any percentile still nearer the extreme of the total group, more pupils are required to secure stable percentile norms than to place a test on an age scale, while the average norm is stabilized by a relatively small number of cases.

Test norms are more useful when the method of their derivation is clearly described. This appears self-evident, yet it is not at all uncommon to find a statement of norms without any explanation as to the method of their derivation, whether, for example, they are mean norms or median norms, or 75 percentile norms, or some other kind.

Test norms are more useful when they are reported in full. One author reports as norms for his test the highest score ever made in the test by any one class in the grade. This may have been done to stimulate teachers to special effort to bring their class up to this high standard. But such a stimulation is as liable to be unwholesome as beneficial since it may lead to overemphasis upon one subject. It is well to give the highest score and lowest score, or better the upper quartile and lower quartile scores, or, better still, all the percentile scores, for the fuller the norms are stated, the better. But whether several norms are given or only one norm, the best single score to report is some average measure.

Test norms are more useful when they are both universal and local. An average norm for wide areas is useful but so are separate norms for a great many typical locations. Williams in North Carolina, Foote in Louisiana, Jordan in Arkansas, and other workers in southern states found that the spread of educational measurement in the South was handicapped by lack of norms from localities comparable with rural schools in southern states. They have begun the work of securing norms for the more important educational tests. While universality and locality of standardization are important considerations, it should not be forgotten that ex-

cellent norms will not make an otherwise poor test into a good one.

Finally, norms are more valuable when reported for both age and grade. Ballard, in his recent book on mental tests, complains that most of the norms developed in America are useless in England because our norms are grade norms. Age norms would be almost as valuable in England as in America. Again, age norms permit the computation of reading age and Reading Quotient, spelling age and Spelling Quotient, and mental age and Intelligence Quotient. Numerous instances of the important functions which such measures serve have been illustrated many times throughout this book.

SUPPLEMENTARY READING FOR PART II

Buckingham, B. R.—*Spelling Ability; Its Measurement and Distribution;* Bureau of Publication, Teachers College, New York, 1913.

Courtis, Stuart A.—*The Gary Public Schools: Measurement of Classroom Products;* General Education Board, New York, 1919.

Dewey, Evelyn; Child, Emily; and Ruml, Beardsley.—*Methods and Results of Testing School Children;* E. P. Dutton & Company, New York, 1920.

Hillegas, Milo B.—*Scale for the Measurement of Quality in English Composition by Young People;* Teachers College, Columbia University, New York, 1912.

Pintner, Rudolf.—*The Mental Survey;* D. Appleton & Company, New York, 1918.

Pintner, Rudolf, and Paterson, Donald.—*A Scale of Performance Tests;* D. Appleton & Company, New York, 1917.

Rugg, Harold O.—*Application of Statistical Methods to Education;* Houghton Mifflin Company, New York, 1916.

Terman, Lewis M.—*The Measurement of Intelligence;* Houghton Mifflin Company, New York, 1916.

THORNDIKE, EDWARD L.—*Introduction to the Theory of Mental and Social Measurements;* Teachers College, Columbia University, New York, 1913.

TRABUE, M. R.—*Completion Test Language Scales;* Teachers College, Columbia University, New York, 1915.

VAN WAGENEN, M. J.—*Historical Information and Judgment of Elementary School Pupils;* Teachers College, Columbia University, New York, 1919.

WOODY, CLIFFORD.—*Measurements of Some Achievements in Arithmetic;* Teachers College, Columbia University, New York, 1916.

YERKES, R. M.; BRIDGES, J. W.; and HARDWICK, ROSE S.—*A Point Scale for Measuring Mental Ability;* Warwick & York, Baltimore, 1915.

YOAKUM, CLARENCE S., and YERKES, R. M.—*Army Mental Tests;* Henry Holt & Company, New York, 1920.

PART THREE

TABULAR, GRAPHIC, AND STATISTICAL METHODS

CHAPTER XII

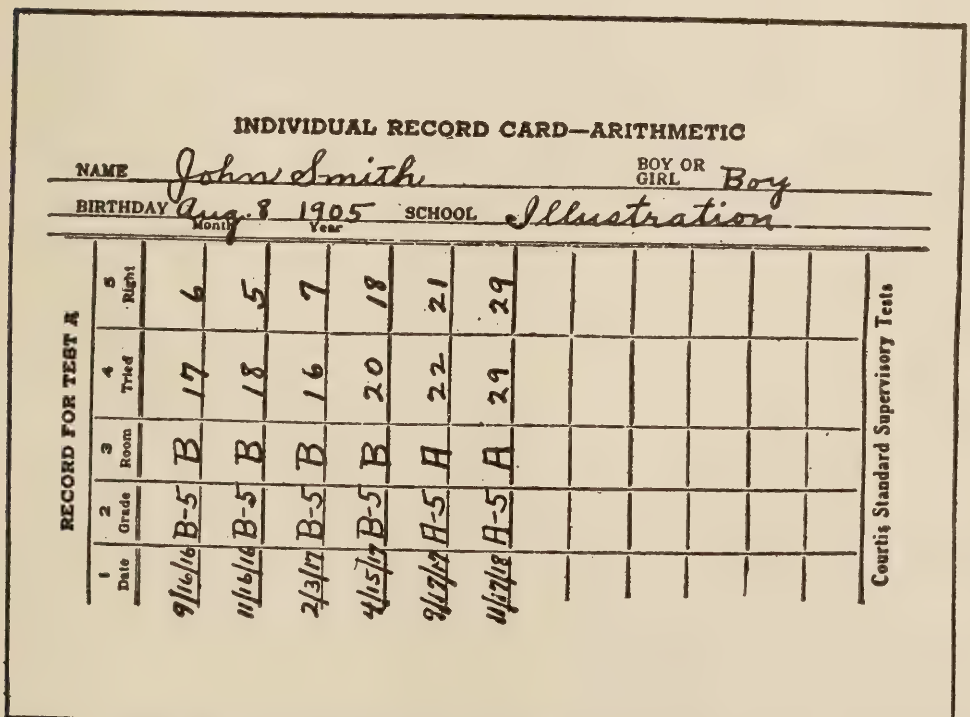
TABULAR METHODS

Types of Tabulation.—Numerous practical and scientific uses of test scores require that they be preserved in convenient tabular form. The types of tabulation are too numerous to describe in detail. The more commonly used varieties are listed below.

1. A tabulation showing one pupil and one test which is non-cumulative. Thus:

Name	Test Score
Adams, Frank	26

2. A cumulative tabulation showing one pupil and one test. Thus:



INDIVIDUAL RECORD CARD—ARITHMETIC

NAME John Smith BOY OR GIRL Boy

BIRTHDAY Aug. 8 1905 (Month, Year) SCHOOL Illustration

RECORD FOR TEST #

1 Date	2 Grade	3 Room	4 Tried	5 Right
9/16/16	B-5	B	17	6
11/16/16	B-5	B	18	5
2/3/17	B-5	B	16	7
4/15/17	B-5	B	20	18
9/17/17	A-5	A	22	21
11/17/18	A-5	A	29	29

Courtis Standard Supervisory Tests

FIG. 4. Shows How to Fill Out the Individual Record Card for One of Courtis' Standard Supervisory Tests.

3. A tabulation showing one pupil and one test which is tabulated by test elements. Thus:

Name	a	b	c	d	e	f	g	Test Score
	Test Elements							
Adams, Frank	1	1	1	0	0	x	x	3

4. A tabulation showing one pupil and many tests. Thus:

Name	Test I	Test II	Test III	Test IV
Adams, Frank	8	72	41	20

5. A cumulative tabulation showing one pupil and many tests. All are familiar with the cumulative record card for teachers' marks for an individual pupil. A cumulative record card for test scores would be similar.

6. A group tabulation showing one test and revealing the identity of each pupil. The addition of several names and scores to No. 1 above would convert it into such a tabulation.

7. A group tabulation showing one test tabulated by test elements and revealing the identity of each pupil. The addition of several names and scores to No. 3 would convert it into such a tabulation.

8. A group tabulation showing several tests and revealing the identity of each pupil. Thus:

Pupil	Test I	Test II	Test III	Test IV	Test V	Test VI
a	13	32	18	42	60	5.0
b	13	35	20	50	72	6.0
c	10	30	12	30	60	4.5

When the tests are very numerous the tabulations could be made in a quadrilled blank book where the edges of the leaves have been so cut away as to make it unnecessary to rewrite the list of names on each page.

9. A group tabulation showing one test and concealing the identity of each pupil. Many of the tabulation forms sent out with tests provide for the tabulation of the pupil's

score into a frequency distribution (see later chapter) where the pupil's name does not appear.

10. A group tabulation showing several tests tabulated into frequency distributions without concealing the identity of each pupil. Test I and Test II under No. 8 above, which were tabulated according to what Rugg calls the *writing* method, are retabulated below according to the *checking* method.

Pupil	Test I								Test II								
	8	9	10	11	12	13	14	15	24-26	26-28	28-30	30-32	32-34	34-36	36-38	38-40	40-42
a						x							x				
b						x								x			
c			x									x					
Total	0	0	1	0	0	2	0	0	0	0	0	1	1	1	0	0	0

When pupils are numerous statistical computation is facilitated by throwing scores into a frequency distribution. The above method of tabulation yields almost instantaneously a frequency distribution as shown in the total column. This method of tabulation is very advantageous in extensive studies and only in extensive studies.

11. Individual or group tabulation by machinery. Many large school systems have found it economical to install electrically driven machines for tabulating and sorting data. The Hollerith Tabulating Machine can punch on a small card an unbelievably large amount of data under a great variety of heads. The Hollerith Sorting Machine will in a brief time sort a large number of these cards according to any desired item. The machine will, in an additional step, count the cards in any item-group. The Powers machine will tabulate data upon cards in a similar fashion. It has an advantage over the Hollerith machine in that it will sort and count at the same time, thus eliminating one step in the process. Since these mechanical devices can only be rented at a considerable charge they are not appropriate for use on small studies. But where much data is to be handled they are a great economy indeed.

Selection of Tabulation Form.—Which tabulation form to select depends upon the purposes which the data are to serve and the persons for whom they are intended. No form will serve all purposes. The following questions will help to prevent overlooking some vital point in selecting the tabulation form or forms:

1. Will the form make it easy to identify the pupil and his score or scores?
2. Will the form permit the addition of scores from year to year in order that teachers and scientific students of education may study the progress of pupils?
3. Will the form permit the addition of scores from pupils tested late?
4. Does the test require tabulation by test elements before individual or class scores can be computed?
5. Will tabulation by test elements make data more useful to teacher?
6. Will the form economize space and fit existing files?
7. Will the form give a bird's-eye-view of large sections of the data at once?
8. Will the form require names to be written more than once?
9. Will the form easily condense into a frequency distribution?
10. Will the form make it easy to lose data or parts of data?
11. Will the form readily reveal that a portion of data is missing?
12. Will the form, in terms of both tabulation and subsequent uses, economize time?
13. Will the form make it easy to overlook data by the sticking together of cards, or by a misplacement in files?
14. Will the form make filing and refiling in alphabetical order very laborious?
15. Will the form require much mechanical work in the examination of the data?

16. Will the form permit computation on the original tabulation sheet?

17. Will the form permit the tabulation of sub-totals, grand totals, and other summaries on the original tabulation sheet?

18. Will the form facilitate the rearrangement of data in all desired orders without recopying?

19. Will the form make it possible to separate any part of the data from other parts so that different individuals may be working upon the data at one time?

20. Will the form stand the necessary wear and tear?

21. Will the spaces on the form fit the requirements of the test or tests now and in the future?

22. Will identification data such as name, sex, birthday, etc., have to be written more than once?

23. Will the form permit the recording of interpretative scores such as quotients?

Construction and Placement of Tables.—Experience has gradually crystallized into a series of conventions concerning how tables should be constructed. These conventions rest upon no particular authority, and are in fact frequently violated either from ignorance of their existence or to save space or for some other temporarily valid reason. A few of these principles are illustrated in Table 28 and are summarized in the following principles:

1. *The table number or letter and the title are placed above the table.* Later it will be noted that the situation is just the reverse for diagrams.

2. *The title is sufficiently clear and complete to make it unnecessary to peruse the accompanying text in order to be able to understand the table.* Along with a table an author usually has a rather full interpretation of it. All of this cannot go into the caption for the table. The title would be too bulky. It is necessary to choose for this caption the most essential features in the light of the most probable uses of the tabular data, and to state this in as concise and yet

TABLE 28

Median Scores Made May, 1917, on Woody's Four Fundamentals of Arithmetic Scales, Series B, Compared With Woody's Norms

School and Test	Grade and Section							
	III		IV		V		VI	
	A	B	A	B	A	B	A	B
N. Y. C. School								
Addition	12.0	12.6	13.3	14.5	...	...	17.5	18.1
Subtraction ..	8.0	9.1	10.0	10.7	...	...	13.7	14.2
Multiplication	7.5	8.2	9.4	11.3	...	...	16.0	17.1
Division	2.1	6.0	7.2	8.0	...	...	11.3	12.4
June Norm								
Addition	...	9.0	...	11.0	...	14.0	...	16.0
Subtraction..	...	6.0	...	8.0	...	10.0	...	12.0
Multiplication	...	3.5	...	7.0	...	11.0	...	15.0
Division	...	3.0	...	5.0	...	7.0	...	10.0
N. Y. C. School Total	29.6	35.9	39.9	44.5	...	...	58.5	61.8
June Norm Total	...	21.5	...	31.0	...	42.0	...	53.0

clear form as possible. The following question tests the excellence of a caption: Could the table be read and understood if it were completely removed from the text?

This does not mean that accompanying text should be eliminated as useless. Alexander [1] is right when he insists that statistical material needs to be *translated.* He offers the following seven suggestions for good translations and profusely illustrates each.

a. The illustrations and images used must be of an elementary nature, or at least familiar to the people for whom the translation is being made.

b. In some cases it may be necessary to use several illustrations in order to be sure of reaching all classes of people.

c. Instead of representing a total by imagining an unreal extension of a familiar object, or by making up from

[1] Carter Alexander, *School Statistics and Publicity;* Silver, Burdett & Company, 1919, 332 pp.

familiar units an aggregate so large as to be incomprehensible, it is usually better to employ some other unit. Often this other unit is one of time.

d. In cost statistics, it is sometimes advisable to minimize the total by expressing it in amount per small unit of time, usually a trivial sum.

e. Absolute accuracy frequently has to be sacrificed to force and clearness in translation.

f. Practically all totals have to be translated through comparisons, using familiar objects or notions, before they can be understood or have much force for the average man.

g. Many questions involving value, and particularly exhibits of loss or waste, can be profitably translated into a money equivalent. This is particularly true of all proposals involving an increase in school taxes, which must, of course, be addressed to the taxpayer.

3. *The descriptive items, such as the name of the school, the tests, the grades and the sections are placed at the left and top.* In the preceding table the grades and sections were placed at the top and other descriptive items at the left. The grades could have been placed at the left and other descriptive items at the top, and under certain circumstances this would be preferable. In this case, however, such an arrangement would be less satisfactory. As the table now stands every descriptive item can be read from the bottom. Were the other descriptive items placed at the top they would extend beyond the limits of the page or they would have to be printed vertically, which makes reading difficult.

4. *Descriptive items are so placed that they may be read from the bottom of the page or bottom and right of the page.* When possible, as it was in the above table, all items should be so placed that they may be read from one point, namely, the bottom of the page. The requirements of space sometimes make it necessary to print the items at the top vertically. When this is done they should be made to read from the right of the page, while those placed at the left should be made to read from the bottom.

5. *Items and data are appropriately grouped and the*

grouping is clearly indicated by indentation, spaces, or lines. Thus "Addition," "Subtraction," etc., in the above table are placed under "N. Y. C. School" and indented. The data for "N. Y. C. School" is set off from "June Norm" by a horizontal space. Sections A and B are placed under the grades and their subordinate nature is still further indicated by means of lines.

6. *Sub-items are placed under super-items and are more indented.* Thus "Addition," "Subtraction," etc., are placed under and to the right of "N. Y. C. School."

7. *The table reads downward and to the right.* This is in part the reverse of diagrams which read upward and to the right.

8. *There are enough lines and rows of dots to guide the eye in reading the table.* Lines and dots not only guide the eye but give the entire table a neater and more artistic appearance. More horizontal guide lines for the above table would be a dubious improvement. Too many lines are as bad as too few.

9. *Important lines are either made double or extra heavy.*

10. *When a table has long columns of figures each group of about five figures is separated from the adjoining group by a space.* This practice is advisable even though the table is not intended for publication.

11. *The print is sufficiently large to prevent eye-strain.* This principle is particularly pertinent in connection with summary tables placed in the body of the text where they will be frequently used. Tables of original data placed in the appendix may legitimately have a finer print.

12. *A gap in the consecutive units is usually indicated by a break in the data.* The above table shows very clearly that Grade V was not tested. This prevents one from drawing the erroneous conclusion from a hasty inspection of the table that there is a sudden spurt in pupils' ability in the fundamentals of arithmetic.

13. *Condensation of a column of figures into a total or*

average is clearly marked off by a line or space. The reason for this is so evident as to require no explanation.

14. *All decimal points in each column are kept in line.* This appears such a truism as not to require mentioning. The statistician is rare, however, who has not spent considerable time recopying tables tabulated by presumably competent adults in such a way that not only were decimal points out of alignment but the whole column wound and twisted down even a quadrilled paper.

15. *When circumstances permit the data are more effective when arranged in an order distribution.* The first column of Table 29 (heavy print not in the original) illustrates this principle.

TABLE 29

Shows the Median Sizes of Classes in a Number of High Schools by Subjects Together with the Range of the Middle Fifty Per Cent (After Bobbitt[2])

Subject	Median No. Pupils	"Zone of Safety"
		Pupils
Music	58	42–88
Physical Training	32	28–55
English	22	20–24
Mathematics	21	18–24
History	21	17–23
Science	20	16–22
Agriculture	19	18–25
Commercial	19	15–23
Drawing	18	14–24
Modern Languages	17	15–20
Latin	17	14–19
Household Occupations.	17	13–23
Normal Training	15	10–21
Shopwork	14	12–18

16. *Important items are made prominent by the use of heavy type.* Let us suppose that Bobbitt constructed the above table to show the number of pupils in English classes

[2] J. F. Bobbitt, "High School Costs"; *School Review,* 23: 505-534.

as compared with the number in other classes. The heavy type focuses attention upon English and the order distribution shows that it holds third position among subjects.

17. *The table is placed as near as possible to the text which interprets it.* Other things being equal, this is the case. Sometimes the table is placed at the beginning of the descriptive text, sometimes at the end. Table 28 was placed at the beginning of these descriptive principles so the reader would have in mind a concrete illustration of most of them. To present a fact and then explain it is psychologically easier to follow than to keep the reader in the dark until the culminating moment of the explanatory process has arrived. Sometimes, when the descriptive text is not dependent for meaning or concreteness upon the table, or when the descriptive text must precede to give any meaning to the table, the table is placed at the end or in the midst of the descriptive text. Sometimes, however, tables are and should be placed in the appendix only. Long tables of original data such as are found in Ph.D. theses would, if placed in the body of the book, tend to terrify the most hardened reader of statistical literature. They rarely illuminate the meaning of the text, and even when necessary for this purpose, a sampling is usually sufficient. Summary tables, which are really meant to be studied, should, however, be imbedded in the text.

The above principle is a protest against the tendency of some authors to put all their tables at one place, which is frequently far removed from the place where these tables are discussed.

Additional suggestions for the construction of tables may be gleaned from the principles for graphic presentation which follow.

CHAPTER XIII

GRAPHIC METHODS

Importance of Presentation.—Other chapters may exceed this one in length, but none exceeds it in the importance of the topic considered. Recently posters appeared on New York City bill boards announcing a new play: "It Pays to Advertise." The poster showed a cackling hen leaving an egg-filled nest. For the sake of the public it is necessary to have a dignified title for this chapter. But it will not be amiss to imbed here in the privacy of the text the statement that the real title of this chapter is: It Pays to Advertise. Preceding chapters have attempted to show how the truth about conditions in the school may be discovered. Presumably these facts have not been collected to fill up files, but rather to publish in the schoolroom, at teachers' meetings, in public addresses, in school reports, or in periodicals. Presumably these facts have been collected to influence action—the action of pupils, teachers, supervisors, principals, superintendents, boards of education, or the public. Truth does not prevail through facts but through the effective presentation of facts.

There are three types of presentation in common use: the tabular, the graphic, and the linguistic. Generally speaking, that type of presentation is most significant which in the particular situation best fits the data, the purpose, the occasion, the medium of presentation, whether in an address, a published article, etc., and which best fits the kind of audience.

The graphic method is, however, generally conceded to be the best method for most situations. The graphic method is particularly effective because when graphs are properly

made they are more easily and more quickly interpreted. For both these reasons, and perhaps others in addition, graphs have an intrinsic psychological appeal denied to numbers and words. It is only the unusual person whose tabular or literary skill is sufficient to overcome this inherent superiority of the graphic method. Finally, the properly constructed graph shows not only the graph but presents tabular data and utilizes linguistic description at the same time. The graph combines most of the advantages of all three methods, and is hence a powerful instrument in the hands of intelligent educators.

Standard Graphic Methods.—The standardization of graphic methods is just as important as the standardization of statistical procedure. In order to further a notable movement toward standardization which has already begun and in order to give the reader an introduction to graphic methods the full preliminary report of the Joint Committee on Standards for Graphic Presentation is given below.

Joint Committee on Standards for Graphic Presentation

Preliminary Report Published for the Purpose of Inviting Suggestions for the Benefit of the Committee

As a result of invitations extended by The American Society of Mechanical Engineers, a number of associations of national scope have appointed representatives on a Joint Committee on Standards for Graphic Presentation. Below are the names of the members of the committee and of the associations which have coöperated in its formation.

Willard C. Brinton, *Chairman,* American Society of Mechanical Engineers.
7 East 42d Street, New York City.
Leonard P. Ayres, *Secretary*, American Statistical Association.
130 East 22d Street, New York City.
N. A. Carle, American Institute of Electrical Engineers.
Robert E. Chaddock, American Association for the Advancement of Science.
Frederick A. Cleveland, American Academy of Political and Social Science.
H. E. Crampton, American Genetic Association.
Walter S. Gifford, American Economic Association.

J. Arthur Harris, American Society of Naturalists.
H. E. Hawkes, American Mathematical Society.
Joseph A. Hill, United States Census Bureau.
Henry D. Hubbard, United States Bureau of Standards.
Robert H. Montgomery, American Association of Public Accountants.
Henry H. Norris, Society for the Promotion of Engineering Education.
Alexander Smith, American Chemical Society.
Judd Stewart, American Institute of Mining Engineers.
Wendell M. Strong, Actuarial Society of America.
Edward L. Thorndike, American Psychological Association.

The committee is making a study of the methods used in different fields of endeavor for presenting statistical and quantitative data in graphic form. As civilization advances there is being brought to the attention of the average individual a constantly increasing volume of comparative figures and general data of a scientific, technical and statistical nature. The graphic method permits the presentation of such figures and data with a great saving of time and also with more clearness than would otherwise be obtained. If simple and convenient standards can be found and made generally known, there will be possible a more universal use of graphic methods with a consequent gain to mankind because of the greater speed and accuracy with which complex information may be imparted and interpreted.

THE FOLLOWING ARE SUGGESTIONS WHICH THE COMMITTEE HAS THUS FAR CONSIDERED AS REPRESENTING THE MORE GENERALLY APPLICABLE PRINCIPLES OF ELEMENTARY GRAPHIC PRESENTATION

1. The general arrangement of a diagram should proceed from left to right.

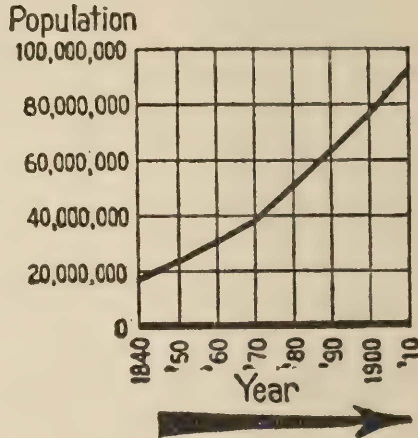


FIG. 5

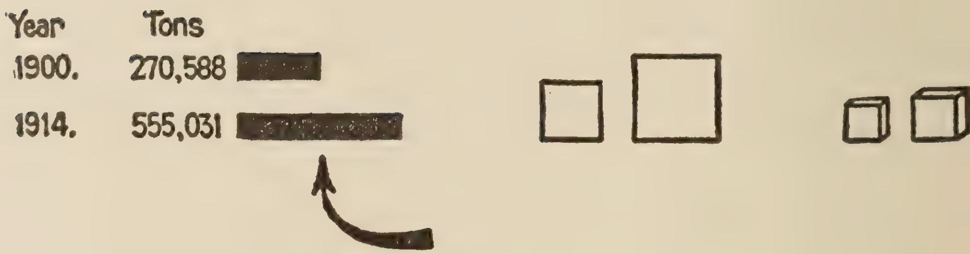


FIG. 6

2. Where possible represent quantities by linear magnitudes as areas or volumes are more likely to be misinterpreted.

3. For a curve the vertical scale, whenever practicable, should be so selected that the zero line will appear on the diagram.

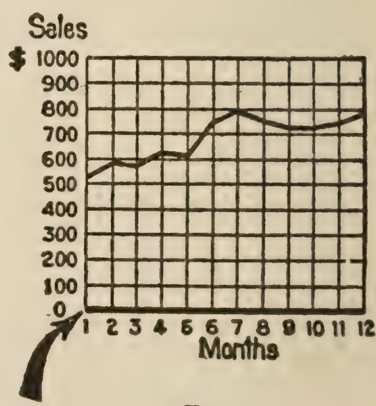


FIG. 7

4. If the zero line of the vertical scale will not normally appear on the curve diagram, the zero line should be shown by the use of a horizontal break in the diagram.

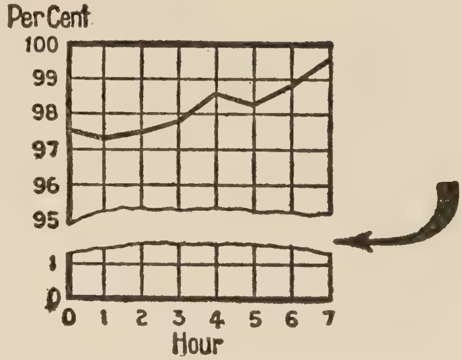


Fig. 8

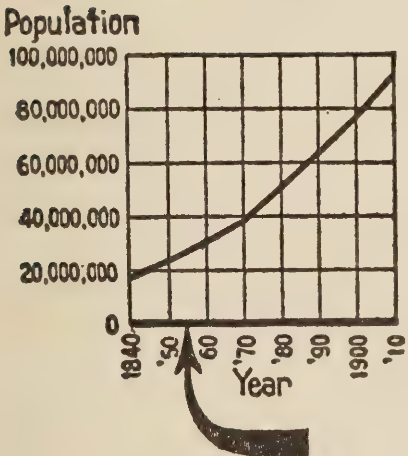


Fig. 9A

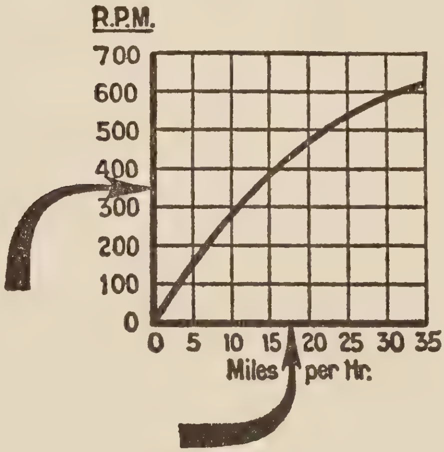


Fig. 9B

5. The zero lines of the scales for a curve should be sharply distinguished from the other coördinate lines.

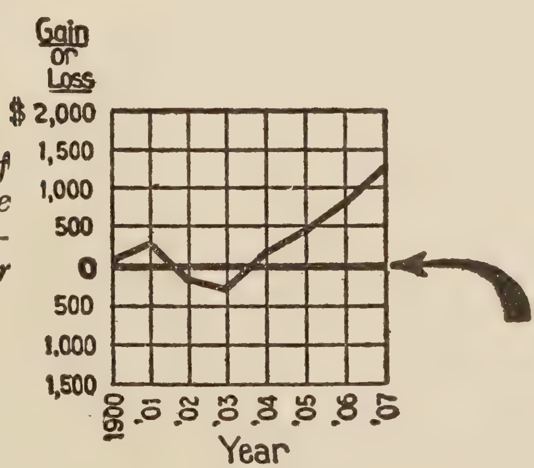


Fig. 9C

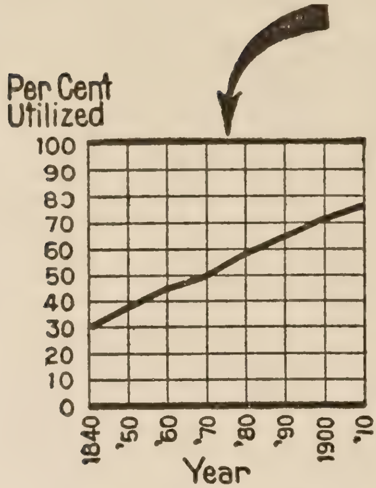


FIG. 10A

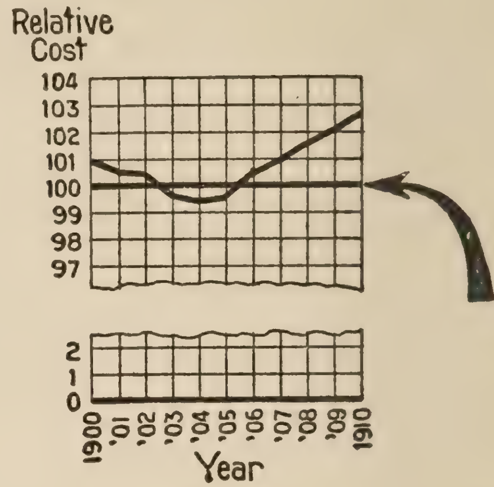


FIG. 10B

6. For curves having a scale representing percentages, it is usually desirable to emphasize in some distinctive way the 100 per cent line or other line used as a basis of comparison.

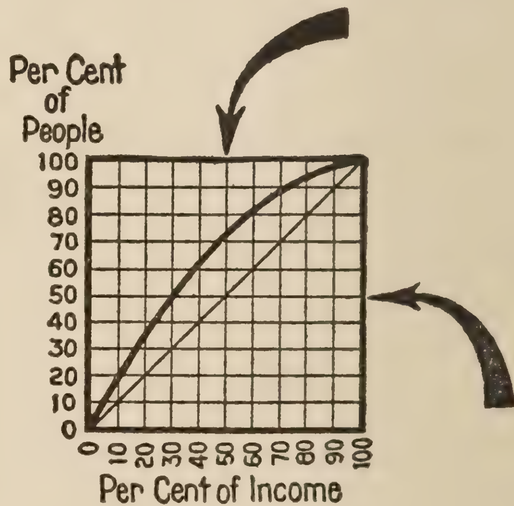


FIG. 10C

7. When the scale of a diagram refers to dates, and the period represented is not a complete unit, it is better not to emphasize the first and last ordinates, since such a diagram does not represent the beginning or end of time.

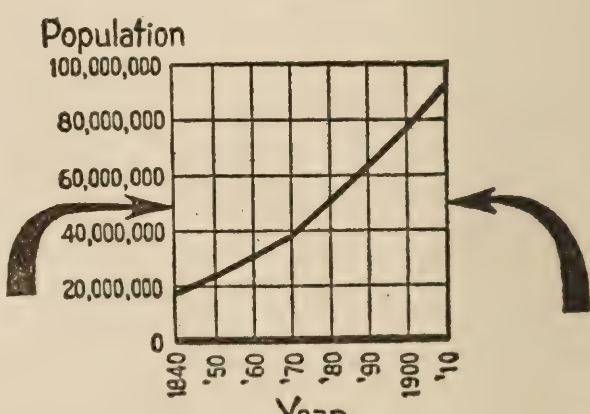


FIG. 11

8. When curves are drawn on logarithmic coördinates, the limiting lines of the diagram should each be at some power of ten on the logarithmic scales.

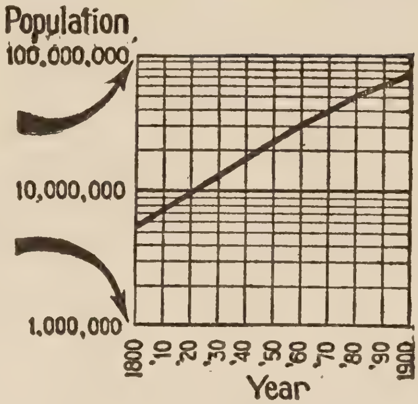


Fig. 12

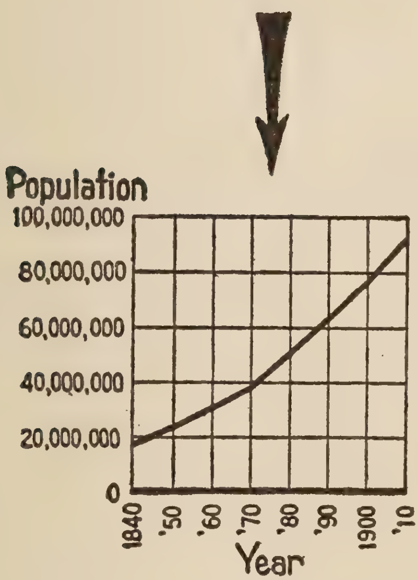


Fig. 13A

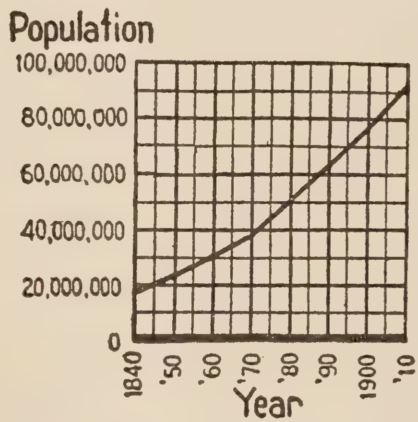


Fig. 13B

9. It is advisable not to show any more coördinate lines than necessary to guide the eye in reading the diagram.

10. The curve lines of a diagram should be sharply distinguished from the ruling.

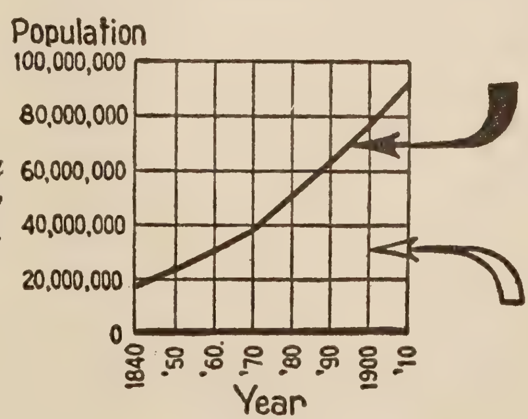


Fig. 14

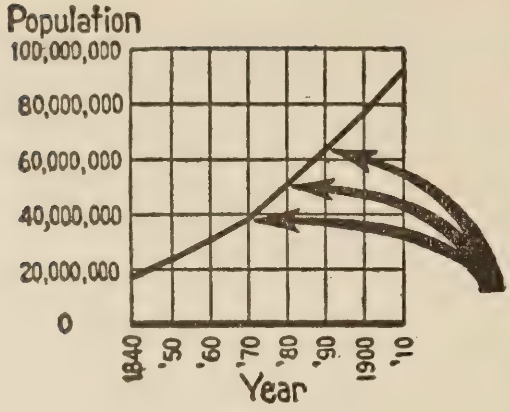


Fig. 15A

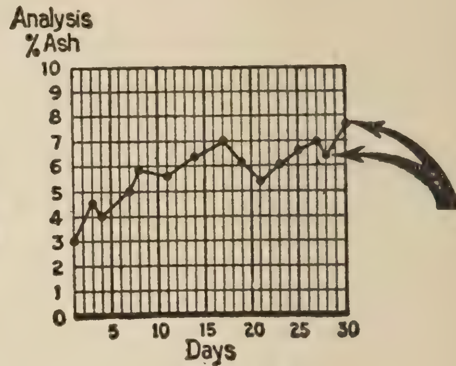


Fig. 15B

11. In curves representing a series of observations, it is advisable, whenever possible, to indicate clearly on the diagram all the curves representing the separate observations.

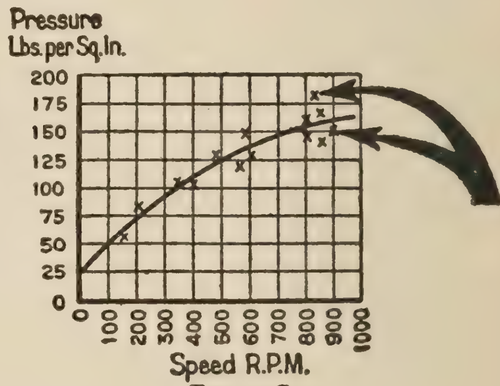


Fig. 15C

12. The horizontal scale for curves should usually read from left to right and the vertical scale from bottom to top.

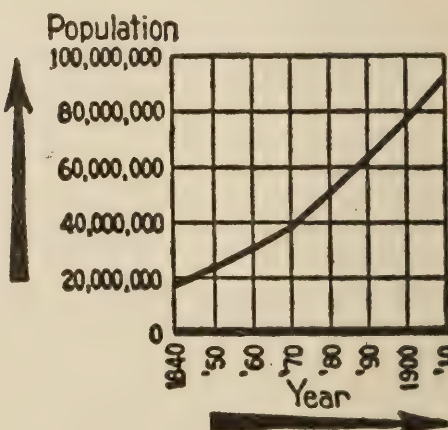


Fig. 16

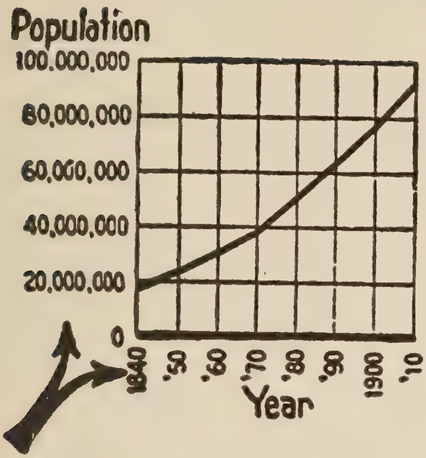


Fig. 17A

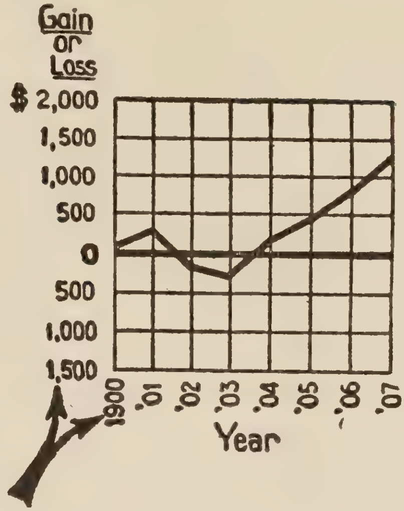


Fig. 17B

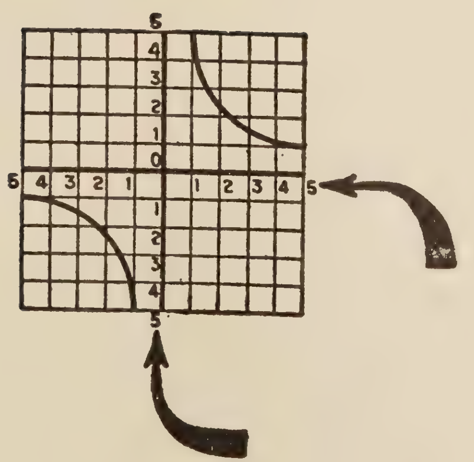


Fig. 17C

13. Figures for the scales of a diagram should be placed at the left and at the bottom or along the respective axes.

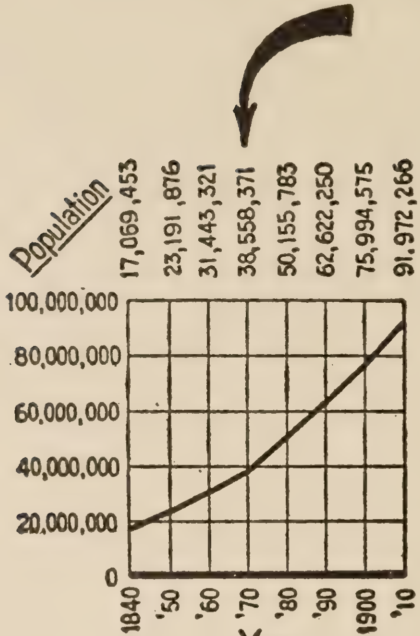


FIG. 18A

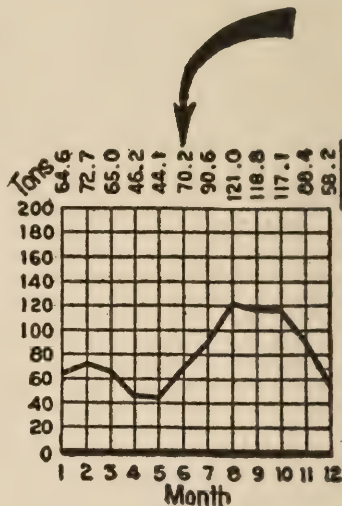


FIG. 18B

14. It is often desirable to include in the diagram the numerical data or formulæ represented.

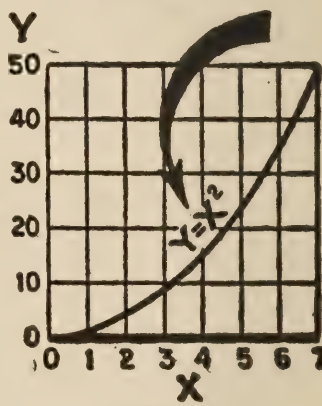
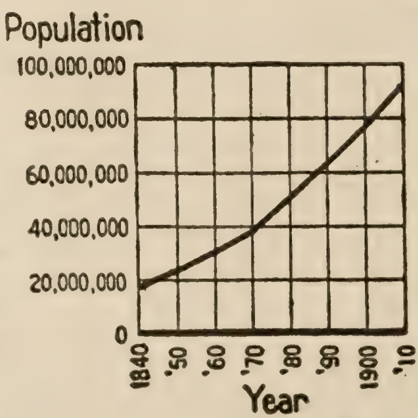


FIG. 18C

15. If numerical data are not included in the diagram it is desirable to give the data in tabular form accompanying the diagram.



Year	Population
1840	17,069,453
1850	23,191,876
1860	31,443,321
1870	38,558,371
1880	50,155,783
1890	62,622,250
1900	75,994,575
1910	91,972,266

FIG. 19

16. All lettering and all figures on a diagram should be placed so as to be easily read from the base as the bottom, or from the right-hand edge of the diagram as the bottom.

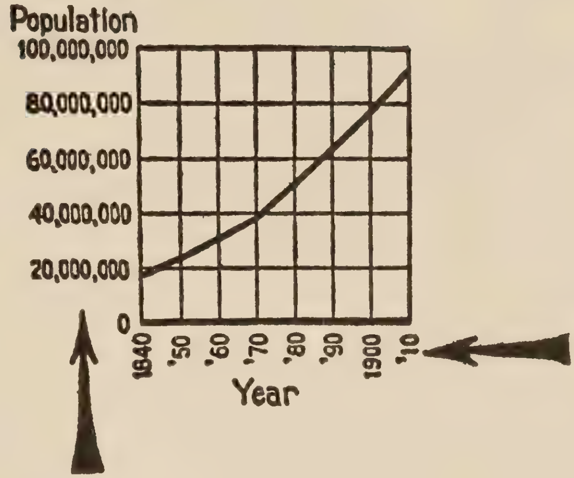
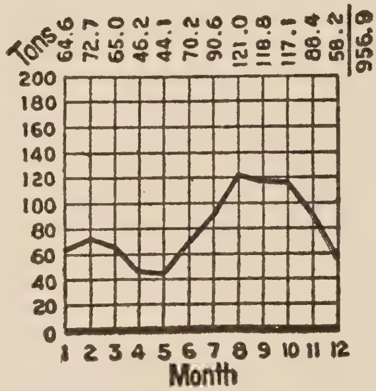


FIG. 20

17. The title of a diagram should be made as clear and complete as possible. Sub-titles or descriptions should be added if necessary to insure clearness.



Aluminum Castings Output of Plant No. 2, by Months, 1914. Output is given in short tons. Sales of Scrap Aluminum are not included.

FIG. 21

Further Principles of Graphing. — The suggestions given below do not appear in the report of the Committee on Graphic Presentation, but through the influence of Brinton's book,[1] in particular, they have become rather generally accepted as good practice. The reader is referred to his book for a further amplification and illustration of these principles.

[1] Willard C. Brinton, "Graphic Methods for Presenting Facts"; *The Engineering Magazine Co.*, New York, 1917, 371 pp.

18. *When several items are being compared the item of chief interest may be made more striking than the others.*

The most important item can be made more striking by the use of (*a*) capitals or red letters for the title. Thus in Fig. 6, for example, the "1914" and the "555,031" could have been printed in red, provided the year 1914 had some peculiar importance. If a principal were comparing his school with other schools he would make the title of the bar representing his own school red, or capitalize the title of his school. If, on the other hand, several schools are being compared with standard, the standard would be made red because the standard would be the most prominent item.

The important item could be made more striking by the use of (*b*) a solid bar for the important item and an outlined bar for the secondary items, or by the use of (*c*) a heavier bar or curve for the important item, or by the use of (*d*) a colored bar or curve for the important item. If desirable and undesirable items are being compared and more than one color is used, it has become a practice to represent the undesirable items by red and the desirable items by green.

19. *Popular features or "eye catchers" may be used to attract attention to the diagram but may not, as a rule, be an integral part of the diagram.*

If the diagram concerns the cost of producing a given unit of growth in pupils large $'s will help to attract attention, but they should accompany the diagram and not be a part of it. That is, no attempt should be made to show the cost by the number of $'s.

20. *Do not place captions or numbers so as to alter the length of bars or to interfere with a visual comparison of their length.*

This means that all numbers should appear at the left of the bars, unless the bars are drawn vertically, in which case the numbers may appear at the top of the bars written horizontally.

Were the numbers shown to the right of the bars in Fig.

6 instead of at their left and were the tons for the lower bar a million or more the 1914 bar would be made to appear longer than it really is, due to the longer length of the numbers representing tons. The caption for each bar could also be so placed as to produce a like illusion.

21. *When a scale (time scale especially) is not consecutive indicate the gap by a wider-than-usual space interval.*

Suppose there were a column of five bars like those of Fig. 6, the top one showing the score made on a test by Grade III and the bottom one showing the score made by Grade VIII. Suppose further that there is no score or bar for Grade VII. The omission of Grade VII should be indicated by a relatively wide gap between the sixth and eighth grade bars. Otherwise the reader is likely to be misled into thinking there is a point in the elementary school where there is an exceptionally rapid growth.

22. *In graphing two or more bars or curves for comparison make their zero lines coincide.*

Anyone who has ever drawn straws to determine who shall get the only apple, or pay for the drinks, knows that he must be suspicious of the apparent length of the straws. We are never sure of our comparison until we discover the zero point of each straw. It is necessary to be equally suspicious of graphs whose zero points are not clearly revealed.

23. *Do not use a percentage curve when it is wished to show the actual amounts of increase or decrease and do not use an amount curve when it is wished to show the per cents of increase or decrease.*

Either a curve must be drawn on a logarithmic scale in order to show both amounts and per cents of change or else two graphs are required, one to show amount and one to show percentage.

As to comparable scaling, it is well to remember that of two curves plotted to the same scale and whose variability is identical, the upper curve will appear to have larger fluctuations. Statisticians are familiar with the notion that the variability of two sets of data cannot well be compared

until the variability of each has been divided by the average of the data from which each variability was computed. This means that the larger the data is numerically the larger will be the amount of fluctuation, even when the percentage of variation remains constant. When it is wished to compare the fluctuation of two curves on the same graph, one of which represents numerically small amounts and the other numerically large amounts, convert the amount curves into percentage curves and interpret in the light of the original absolute amount of each.

24. *Use a diagram which is appropriate to the data to be presented.*

What diagrams to use in a given situation is discussed below.

Types of Diagrams.—There are a bewildering variety of diagrams, some good, some bad. And there is an unlimited number of graphs which may be classed as cartoons. Such, for example, is a drawing showing which of a pupil's neural pathways are in action when he is adding, or a drawing which pictures the number of germs in the water where pupils swim or any other of the numerous pictographs. The value of such cartoons usually disappears with use and hence they are not appropriate material to consider here. A ride on a street car, or a brief study of bill boards will give enough suggestions of cartoons to use. To standardize them would be to destroy their value.

Most of the standard diagrams are variations upon a few simple types. The few types listed below will be found adequate for most persons and most purposes. If any reader plans to do a great deal of graphing he should consult some special treatise on the subject, such as Brinton's.

Type I. The sector diagram. Thus far in this book no illustration of the sector diagram has been printed. One is given in Fig. 22.

The construction of a sector diagram is exceedingly simple. There are 360 degrees in the circle. Sixty-six per cent of 360 degrees is 237.6 degrees. The 237.6 degrees

may be roughly estimated with the eye or more accurately measured with a protractor. The other sectors are determined in a similar fashion. The diagram would be much more striking if *each* sector were colored to fit the race which the sector represents.

Type II. The bar diagram. See, for illustration, Fig. 6.

Type III. The sectioned-bar diagram, (*a*) without sub-

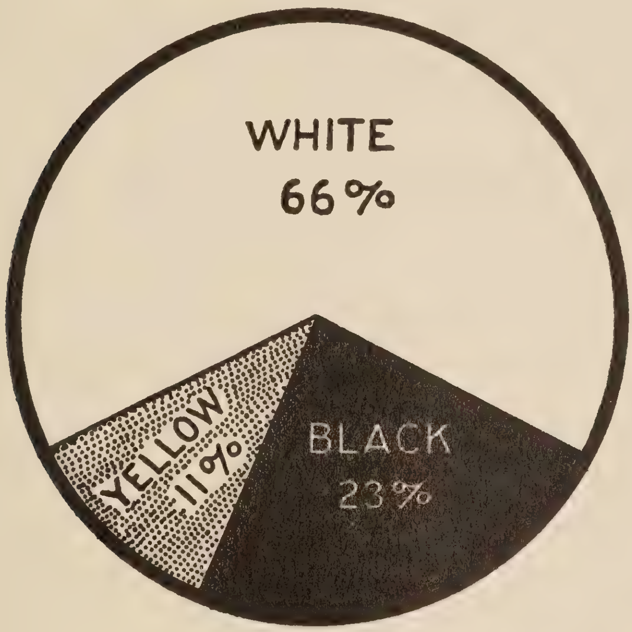


Fig. 22. Distribution by Race of the Pupils in Grades III Through VIII of a Public School in an Eastern City.

divisions, and (*b*) with sub-divisions of the component parts. The top bar of Fig. 23 illustrates the sectioned-bar diagram without sub-divisions, while the entire figure illustrates the diagram with sub-divisions.

This diagram uses such a design in each section as to make it appear distinct from the adjoining sections. This plus the sectioning makes it clear at a glance that in this particular school the per cent of pupils attaining standard gradually increases with progress through the grades. A larger percentage of girls attain standard than boys. With

progress through the grades the percentage of boys who attain standard gradually increases relative to the percentage of girls. In Grade VII the boys' percentage has reached the percentage of the girls.

The unique combination of the bar and sectioned-bar diagrams shown in Fig. 24 is not only unusual but also unusually effective.

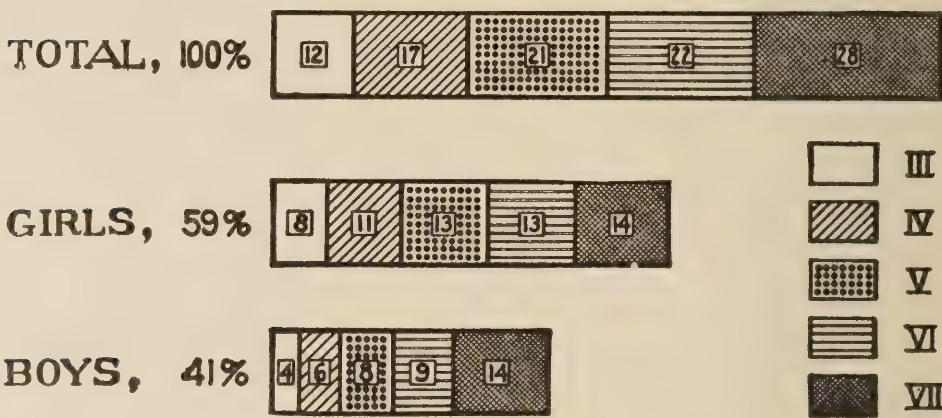


FIG. 23. The Per Cent Which the Number of Pupils in Each Grade Was of the Total Number of Pupils in All Grades Who Attained Woody's Norms According to a Random Sampling of 300 Boys and 300 Girls in a New York School.

Type IV. The frequency surface. Figs. 26 and 28 may be examined as illustrations of frequency surfaces.

Type V. The curve diagram. Numerous illustrations appear in the Report of the Joint Committee on Standards for Presenting Facts. Fig. 5 may be inspected as a sample.

Practically every diagram listed above, except the sector diagram, is a bar diagram or some variation on this basic type. The sectioned-bar diagram is merely a bar diagram divided into component parts. The frequency surface is merely a series of bar diagrams placed close together and in a vertical position. A curve diagram is merely a series of non-adjoining narrow vertical bars which are connected at the tops with a continuous line or curve.

A special form of the curve diagram frequently used by mental measurers is the psychograph or mental profile. If the zero line in Fig. 9C represented the standard scores on

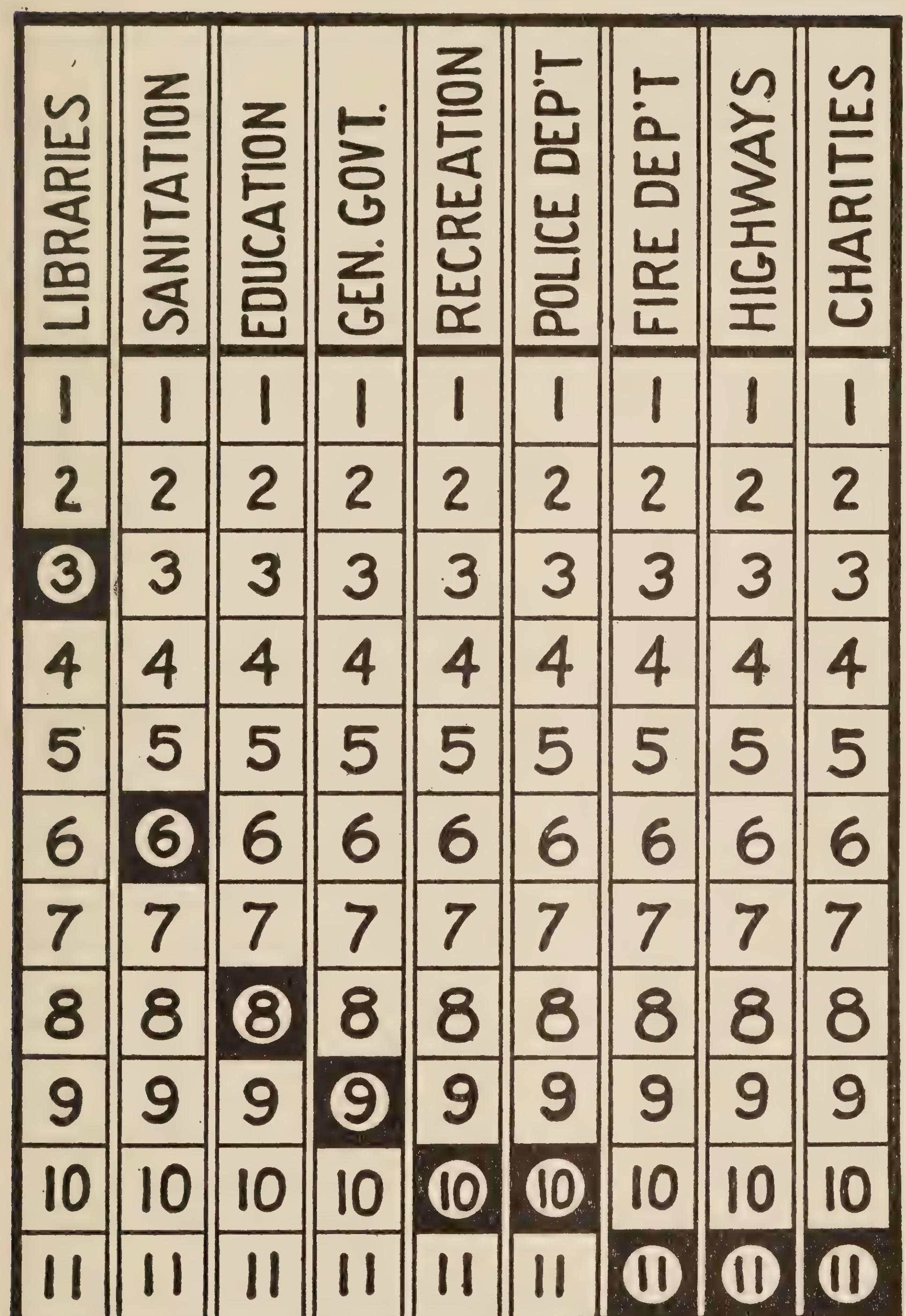


FIG. 24. Rank of Cleveland among Eighteen Cities in Expenditure for Operation and Maintenance of Schools. (After L. P. Ayres, *The Cleveland School Survey*, 1916.)

several mental tests and the dates shown at the bottom were each the name of a mental test, then the second curve would show a sort of mental profile of an individual or group.

Selection of Diagram to Show Component Parts.—Frequently in educational measurement it is necessary to show what part each of various components is of the whole. In order to assist an audience to properly interpret certain test results it may be necessary to show what per cent of the total number of pupils in a school system belongs to the White race, Black race, etc. It may be necessary to show what per cent of pupils in Grade IV of a certain school are eight, nine, ten, or eleven years of age. It may be desirable to show how many or what per cent of the pupils, or schools, or cities make various scores on the test. All these are situations involving component parts of a whole, and require a diagram appropriate to component parts.

Perhaps the simplest of all diagrams for showing component parts is a sector diagram such as is shown in Fig. 22. The sector diagram would serve for any situation listed in the preceding paragraph.

The sectioned-bar diagram shown in Fig. 23 is an even better graph for presenting component parts. It is in almost every respect superior to the sector diagram. Visual comparisons of the components are easier. The direction of all lettering is uniform. The numerical data can be so placed that numbers and decimal points are directly under each other, so that the addition of any or all components is greatly simplified. The sector diagram is not nearly so flexible. It will satisfactorily show only one series of components. The sectioned-bar diagram will show one or more sub-divisions of components. Hence, except in the situation noted below, the sectioned-bar diagram should usually be preferred to other diagrams for showing component parts.

When it is wished to show the number or per cent of pupils making various scores or who are of various ages, or in any situation where the unit is a consecutive numerical fact such as scores, ages, dates, and the like, the frequency

surface is the most convenient graph, although any of the others could be used.

There are several useful variations of the frequency surface. Fig. 25, for example, reveals not only the number of schools making various scores on a test but also the identity of each school making a given score. Again, a frequency surface will show sub-divisions of component parts, in which case the graph really becomes a series of vertically arranged sectioned-bars.

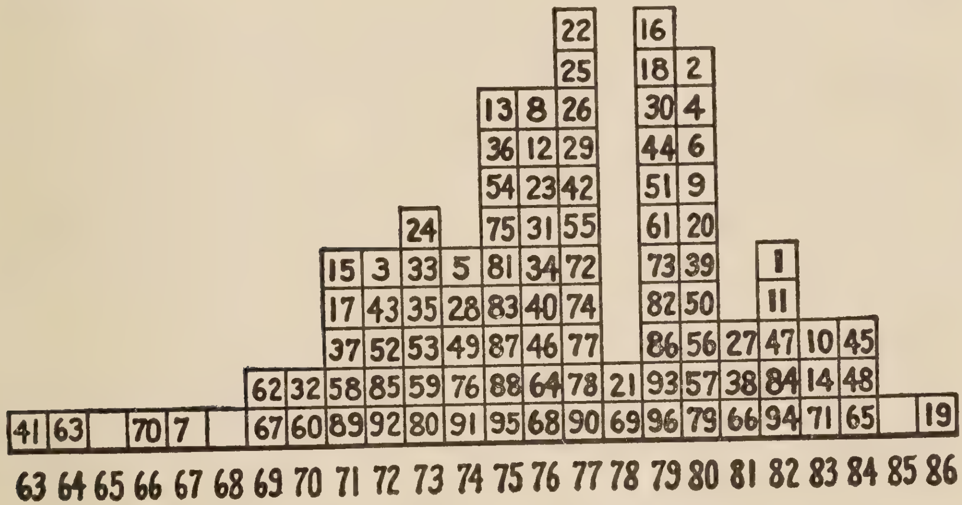


FIG. 25. The Number of Cleveland Elementary Schools and the Identification Number of Each School Making Various Average Scores in Spelling. (After C. H. Judd, *Measuring the Work of the Public Schools,* Russell Sage Foundation, N. Y., 1916.)

Selection of Diagram to Show Comparisons.—For simple comparisons, the best diagram is the bar diagram, an example of which is shown in Fig. 6. The bar diagram can be used to advantage in such situations as the following: where it is necessary to compare (*a*) the number of pupils in one grade with the number of pupils in another grade, (*b*) the norm on a test for, say, Grade III with the median score made by a class in Grade III, (*c*) the score of a grade in one school with the score made by each of several similar grades in other schools, (*d*) the median score made by one grade with the median scores made by each of several grades, (*e*) the score made by one pupil in a class with the

score made by each of the other pupils. No matter how numerous the items, wherever only simple comparisons are involved the bar diagram is thoroughly satisfactory.

Special variations on the diagram can be made to suit special situations. If, for example, the scores of pupils in a class be represented by a series of horizontal bars, one vertical bar can show the median for the class, another can show the norm, thus making possible a comparison of each pupil with every other pupil, with the median for the class, and with the norm.

A bar diagram is not satisfactory for comparing two different series of components. The sector diagram, however, will permit such a comparison. If we were to place beside Fig. 22 another circle of equal size showing similar facts for another school system it would be possible for the eye to roughly compare the sectors. Other graphs, however, permit an easier and more accurate comparison.

The sectioned-bar diagram will show comparisons between two series of data better than the sector diagram. If we were to place one or more graphs, showing similar data, for another school directly under the top bar of Fig. 23 the eye could, with some difficulty, compare the length of one section with the corresponding section. Comparison is made difficult by the fact that the beginning points of all corresponding sections are not directly over each other.

The frequency surface is even more useful than the sector or sectioned-bar diagrams for comparing series of components. Fig. 2 illustrates such an use. Here the frequency surfaces are placed one above the other. When not more than two series of components are being compared the two frequency surfaces may be placed on the identical base line. When there are more than two surfaces on the identical base line the overlapping becomes too confusing to be useful.

The curve diagram is the most useful of all graphs. Its prevalence in the Report of the Joint Committee on Standards for Presenting Facts is a sort of index of its utility. Before finally choosing the type of diagram for presenting

his data the reader will do well to go through the charts of the Joint Committee to see if some curve which he finds there may not satisfy the condition of his data. The curve is familiar to most persons; it is easily and quickly read; it is so flexible that almost any data can be presented by means of it.

The curve is particularly effective for comparing two series of similar data. Suppose we have a curve showing the progress of the medians from grade to grade of a certain school on a certain test. One or more other curves representing the grade progress of other schools on the same test may be drawn on the same diagram, thus permitting easy comparison.

Curve diagrams may also be used to compare series of components. The curve diagram could take the place of the overlapping frequency surfaces in Fig. 2. When a frequency surface is made with a series of rectangles as in Fig. 2 it is called a *histogram*. When a frequency surface is made with a continuous curve it is called a *frequency polygon.* All that is needed to convert the histogram into a frequency polygon is to draw a continuous line which passes through the middle point of the top of each rectangle and then erase the lines which block out the rectangles. In practice, the frequency polygons are drawn directly from the data.

The curve is equally preëminent for showing the relationship between two series of data. A curve of grade progress shows the relationship which obtains between grade and score on a test. A curve of age progress shows the relationship between the age of a pupil and score on a test. Fig. 15C is an illustration of how the curve type of chart may be used to give a graphic picture of correlation.

Preparation of Diagrams.—The following materials are either essential or useful in charting: appropriately ruled paper or plain paper to be ruled by the person making the chart, drawing board, T-square, decimal scale ruler, French curves, reducing glass, colored crayons, waterproof India ink of various colors, gummed letters and figures.

Still other appliances would be useful but few persons outside of professional draftsmen have half the material already mentioned.

The material upon which the diagram is drawn will vary with circumstances. When test records are kept on file from year to year it will be found advisable to make diagrams for these files on ruled cards of uniform filing size. For lecturing purposes the chart may, by means of a brush, be drawn in white paint on black cambric cloth. When the paint has dried this cloth can be folded and packed into a small space in a handbag.

The charts should be drawn in harmony with the suggestions already presented. Besides this, the diagram should be neat with all lettering as plain as possible. Gummed letters and numbers may be used to produce a clearer and neater picture, if the person making the diagram is not skilled in making letters and numbers. If the diagram is intended for publication it is advisable to make the drawing larger than it will be when published, in order that in the process of reduction to printing size minor irregularities will disappear. If the graph is made just twice the printing size great care must be taken to see that every proportion of the original is exactly twice the size that is finally desired. Coördinate and other lines must be twice as wide, and twice as far apart. Letters and numbers must be twice as high and wide and so on. These proportions may be determined by general judgment, by use of the reducing glass, or by actual measurement. Finally, the diagram should be drawn in India ink in order that it may give a clear photograph. Black, red, green, and blue India inks all photograph black. Black prints blacker than any other color; red is a close second, and the others in the order named.

Reproducing the Diagram.—If the diagram is intended for local use it may be reproduced on a hectograph or mimeograph at very little expense and with very little trouble. If this method of reproduction is used the diagram must be prepared with a special kind of ink in the former

case, or on a special stencil in the latter case. Adequate instructions for this process come with these reproducing instruments. A school can ill afford to be without either a hectograph or mimeograph.

The blue print is another method of speedy and inexpensive reproduction and so is the photostat machine. The photostat machine will make direct photographic copies of diagrams. Blue-printing and photostat companies will be found in most large cities.

The stereopticon, reflectoscope, and motion picture may be considered reproducing machines. There are companies who will convert any diagram into a lantern slide whose use in connection with a stereopticon will throw the diagram on a screen. Many schools are finding the stereopticon an indispensable adjunct. There are portable stereopticons which may advantageously be taken on lecture tours. Reflectoscopes are made which will reflect a diagram directly from the paper drawing. This saves time and expense involved in having lantern slides prepared but it is not so satisfactory in other respects as the stereopticon. All are familiar with the motion picture machine.

If a diagram is published one of three methods may be employed, (*a*) a zinc or line cut, (*b*) half-tone or copper plate, or (*c*) Ben Day. The zinc cut is the cheapest, the half-tone next, and the Ben Day process is the most expensive. As stated before, diagrams are usually more effective when printed in color, but color printing is very expensive indeed. Before the diagram is sent to the publisher instructions as to the process and the final dimensions desired should be noted on the margin, preferably in blue pencil since such markings do not photograph. If the process is Ben Day the shading desired for each portion of the graph should be selected from a catalog and indicated.

CHAPTER XIV

STATISTICAL METHODS—MASS MEASURES

Three Types of Mass Measures.—The first step in a statistical study of scores is to convert the data into one or more mass measures. This step is the continental divide between correct and incorrect statistical procedure, and should not be omitted even where it is not necessary to subsequent computation. What statistical measure to compute, whether to compute any measure at all, how to interpret the statistical measures when computed, all three questions depend for their answer in part upon one of the following three mass measures, especially the first or second.

I. Frequency Surface.
II. Frequency Distribution.
III. Order Distribution.

I. Frequency Surfaces

Normal Frequency Surface.—Suppose we have the following table of scores:

TABLE 30

Specially Chosen Scores Made by Sixty-four Fourth-Grade Pupils on a Spelling Test

15	17	14	19	14
11	16	11	15	11
11	9	15	10	19
13	17	18	7	12
9	10	16	16	13
12	14	12	11	7
8	18	17	17	14
10	12	12	14	9
13	15	16	16	14
13	18	13	13	13
10	10	11	20	12
15	13	8	8	6
14	15	12	9	

In their present form these scores are practically uninterpretable, but observe the difference when they are turned into the frequency surface of Fig. 26.

How to Construct the Frequency Surface.—

1. The base line of the graph is drawn.

2. The smallest score in the table is 6 and the largest is 20. Beginning at the left, since the left always means low scores, 6, 7, 8 and so on to 20 are written under successive vertical lines.

3. The first score in the table is 15. Now for a pupil to get a score of 15 in this test means that his true score is somewhere between 15.0 and 15.999, etc. Hence a dot is

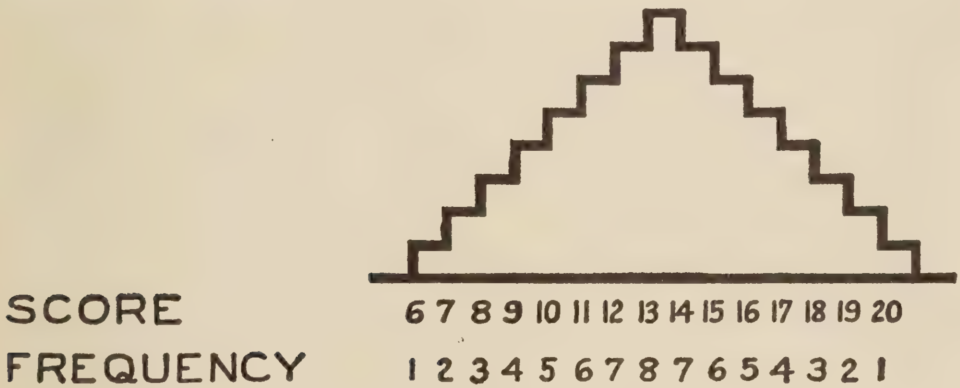


Fig. 26. An Approximately Normal Frequency Surface.
(Data from Table 30.)

placed in the square just above and to the right of 15, i. e., in the square just above the distance 15.0 — 16.0. The next score is 11 and a dot is placed just above and to the right of 11. The next score is also 11, but instead of placing two dots in one square, the dot is placed in the square immediately above the square holding the preceding dot. In this way one square represents one person. This process is continued until every score is checked into its own proper square.

4. The checked squares are traced with a boundary line in block fashion, and we have the frequency surface of Fig. 26.

How to Read the Frequency Surface.—Fig. 26 re-

duces confusion to order and permits one quickly to grasp the total condition of the grade. Besides this, the graph tells the following story almost at a glance. (1) The smallest score is 6 and the largest 20. (2) There are one score of 6, two of 7, three of 8 and so on. (3) The test has been neither too easy nor too difficult, for the scores do not pile up at either the low or high end of the distribution. (4) The variation in scores from 6 to 20 is continuous. (5) The distribution of scores is symmetrical. (6) There is but one central tendency or mode, hence the sur-

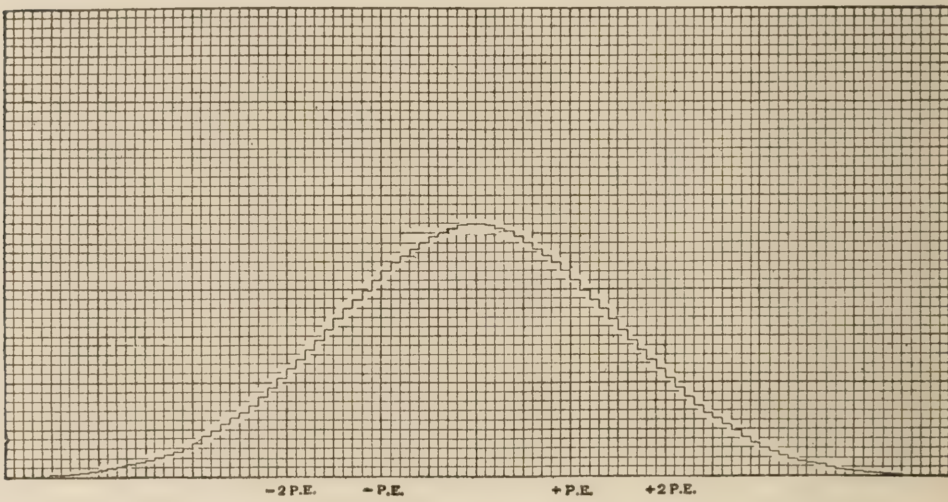


FIG. 27. A Normal Frequency Surface.

face is unimodal, i. e., most of the scores tend to cluster or swarm about one mode or point. (7) The crude mode is at 13 which has a frequency of eight. (8) The frequency surface is a rough approximation to the normal frequency surface, and hence the subsequent statistical methods appropriate to the normal distribution will apply fairly well to the data of Table 30. Fig. 26 only approximates the normal surface for the latter is a smooth curve shaped more like a bell and less like a pyramid, thus giving an even greater clustering toward the central tendency, as may be seen from Fig. 27.

Why Frequency Surfaces Are Normal.—The normal

frequency surface appears to be Nature's favorite mold. A random sampling of most facts gives the normal surface. Morality, intelligence, the weights and heights of men, the blueness of eyes and doubtless the intensity of halos fit the normal curve. It is seldom that mental and educational scores make a *perfectly* normal surface, but they usually give a rough approximation to it. This is, according to Thorndike, not because Nature abhors irregular distributions but because there are usually present in nature the necessary determiners.

Experimental research has isolated these determiners, and has found that the measurements for a given fact fit

Fig. 28. Minus Skewed Frequency Surface.

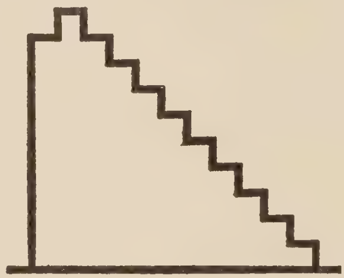


Fig. 29. Plus Skewed Frequency Surface.

the normal frequency surface, when the fact measured is the product of the joint action of, (1) a large number of causes, (2) causes which are approximately equal, (3) causes which are mutually uncorrelated or act independently of each other, i. e., the presence of one cause does not bring with it the other causes. It is because these conditions are usually present in education that most educational measurements approximate the normal distribution.

Skewed Frequency Surface.—One form of departure from the normal surface is called the *skewed* frequency surface. Skewed surfaces may be either minus or plus. Fig. 28 represents a minus skewness and Fig. 29 a plus skewness.

Why Frequency Surfaces Are Skewed.—It is actually

possible to secure either one of these surfaces or approximations to them from the application of a spelling test. Let us enquire why Fig. 28 shows a minus skewness. A complete knowledge of the situation would be necessary before an answer could be given. But it is possible to list some of the more probable causes.

(1) Possibly the test consisted of but sixteen words and all of these were too easy for the abler pupils, thus causing a piling up of the scores at 15 and 16. Ordinarily we should expect the mode to be at 16 instead of 15, but very able pupils often make a score a few less than perfect from purely accidental misspelling.

(2) Possibly the best half of the grade in spelling had been promoted just before the test was administered. In this event we can think of Fig. 28 as being the lower half of a once normal surface.

(3) Possibly the school is located in a neighborhood where the intellectual composition of the parents corresponds to this surface, thus causing by heredity a similar condition among the pupils.

(4) Possibly the native abilities of the pupils are distributed normally and the form of this surface is produced by the teacher's method of ceasing to drill pupils who have attained a standard of 16. A number of other teaching methods would produce similar results.

(5) Possibly the frequency surface is not due to one cause but to the coöperation of two or more of the previously mentioned causes.

It should be observed that all these explanations contemplate a departure from (1) a large number of causes or (2) equality of influence of causes or (3) uncorrelated causes. Most of the explanations have contemplated a very few causes each of which has enormous potency in determining the form of the frequency surface. The reader will find it profitable to stop and enumerate the more probable causations of Fig. 29.

Multi-Modal Frequency Surface.—Another departure

from the normal surface is called the *multi-modal* frequency surface. As the name implies, it has two central tendencies or modes. Such a surface is Fig. 30.

Why Frequency Surfaces Are Multi-Modal.—Multi-modal surfaces may be produced by a number of causes, but most of them are the result of the operation of some one large force of great potency. The most common cause of all multi-modal surfaces is the mixing of two non-homogeneous groups. If, for example, scores from two

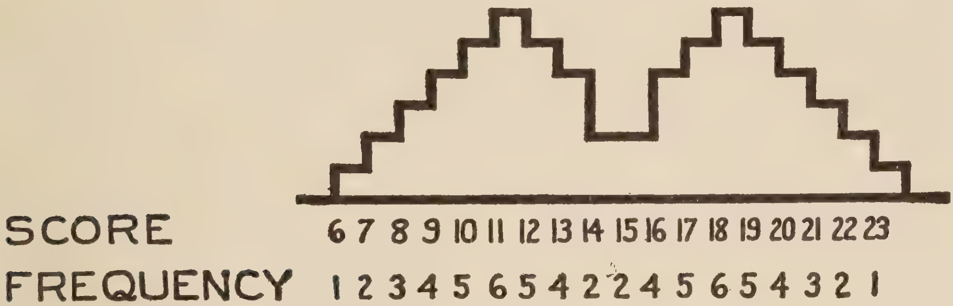


Fig. 30. Multi-modal Frequency Surface.

different grades are graphed together the result is usually a multi-modal frequency surface.

II. Frequency Distributions

Comparison of Frequency Surface and Frequency Distribution.—The second mass measure is called the *frequency distribution,* and, like the frequency surface, appears in three forms, normal, skewed, and multi-modal. In fact, the only difference between a frequency surface and a frequency distribution is this: the former shows in graphic form what the latter shows in tabular form.

The very best method of constructing a frequency distribution is to first construct its corresponding surface. The two rows of numbers beneath the base line of Fig. 26 is an approximately normal frequency distribution. The frequency of score 6, for example, is found by counting the number of squares checked above it, in this case 1, and similarly for the frequency of other scores. Figs. 28, 29,

and 30 show, respectively, at their bases, two skewed and one multi-modal frequency distribution.

Step Interval.—In the illustrations thus far the step interval has been 1. Often the scores are so scattered that to use the original unit of scoring for a step interval would give a frequency surface or distribution with a very much attenuated appearance. This makes interpretation and subsequent manipulation difficult. For practical work Rugg

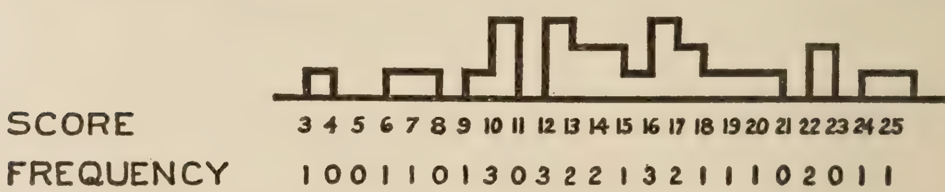


FIG. 31. A Frequency Surface with Step Intervals of 1.

recommends that the step interval be of such size as to make the number of columns in the surface, or the number of steps in the distribution not more than 20 nor less than 10.

Fig. 31 shows scores grouped according to the size of the original scoring unit of 1, while Fig. 32 shows the scores regrouped into a step interval of 2. All the scores of Fig.

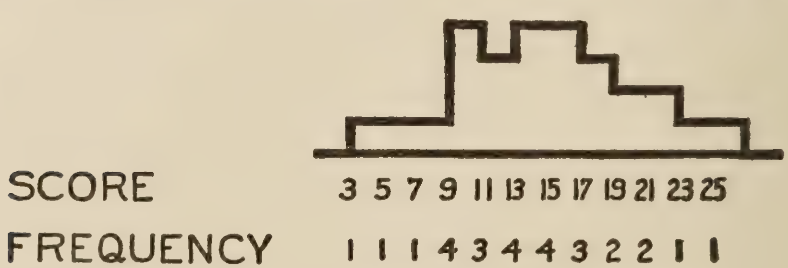


FIG. 32. A Frequency Surface with Step Intervals of 2.
(Data from Fig. 31.)

31 from 3.0 up to but not including 5.0 are regrouped in Fig. 32 over the single interval 3.0 to 5.0. Those from 5 to 7 are grouped over the second interval of Fig. 32 and so on.

In actual practice Fig. 32 is derived immediately from the original data. We do not first plot an attenuated frequency surface and then convert it into a less attenuated

one. It is of course more difficult to fix upon a satisfactory step interval from the original data. The following procedure will make the problem easy: (1) Subtract the smallest score in the original data from the largest. (2) Divide the remainder by such a divisor as will give a quotient between 10 and 20. (3) Make the divisor the size of the step interval. The size of this step interval should be kept constant throughout the frequency surface or distribution.

All the frequency distributions that have been shown thus far have been located beneath frequency surfaces and placed horizontally. In actual computation frequency distributions may be left in this position, but it is usually more convenient to place them in a vertical position. The frequency distribution shown in Fig. 32 is divorced from its frequency surface and placed vertically in the first part of Table 31. It is condensed into larger step intervals in the second part of the table.

TABLE 31

Shows First the Frequency Distribution of Fig. 32, and Second, this Same Frequency Distribution Condensed into Step Intervals of 4

Score	Frequency	Score	Frequency
3—5	1	3—7	2
5—7	1	7—11	5
7—9	1	11—15	7
9—11	4	15—19	7
11—13	3	19—23	4
13—15	4	23—27	2
15—17	4		
17—19	3		
19—21	2		
21—23	2		
23—25	1		
25—27	1		

How to Fix and Indicate the Step Limits.—No statistical computation should be begun until the step limits have been determined. It is not enough to know that the

size of the step interval is .5, 1, 2, 3, or more. We must know from what point to what point the step interval extends. The meaning of the score tells us the beginning point of each step interval and the size of the step interval tells the ending point of each step interval. The meaning of each score in turn depends upon how the test was scored, and the step interval, as we have already seen, depends upon our own choice. Here are some test scores: 6, 7, 8, 9, etc. What is the meaning of each score? We cannot tell without further information as to how the test was scored. Let us suppose that these are scores from some performance test in arithmetic. Most performance tests are so scored that if a pupil works 6 examples correctly he receives a score of 6; if he works 6.25 or 6.5 or 6.75 or 6.9 examples correctly he is still scored 6. Hence in this case 6 means 6 — 6.999, etc., or more conveniently 6 — 7. In other words the beginning point of our first step interval is 6.0.

What is the step interval? We must decide this before the upper limit of the first step interval can be fixed. The step interval cannot in this case be less than 1; it may be made as large as we think desirable. Suppose we make the size of the step interval 1. Then the steps of the frequency distribution would be, 6 — 7, 7 — 8, 8 — 9, etc. If we make the size of the step interval 2, the steps become, 6 — 8, 8 — 10, 10 — 12, etc.

But if our scores of 6, 7, 8, etc., came from a product scale instead of a performance scale, in all probability 6 does not mean 6 — 7, but 5.5 — 6.5, because most product scales, the Thorndike Handwriting Scale, for example, are so scored that a score means half a step below to half a step above a given scale value. If a pupil's handwriting is nearer in quality to value 6 than to any adjoining value he is scored 6, i. e., 5.49 is scored 5, while 5.5, 5.7, 5.9, 6.0, 6.3, 6.4999 or any intermediate values are scored 6. Thus *the meaning of the score depends upon how the test was scored.* Now if we make the size of the step interval 1, the steps will appear as follows: 5.5 — 6.5, 6.5 — 7.5, 7.5 — 8.5,

etc. If we make the size of the step interval 3, the steps become, 5.5 — 8.5, 8.5 — 11.5, etc. Here are some scores from the Ayres Handwriting Scale, 30, 40, 50, etc. The size of the step interval cannot be less than 10 because the values on the Ayres scale are given in units of 10 and because the scorer has obviously not attempted to score between the values appearing on the scale. Though the Ayres scale may be so scored that 30 means 30 — 40, it, being a product scale, is customarily scored in such a way that 30 means 25 — 35. If we make the size of the step interval 10, the steps of the frequency distribution will appear as follows: 25 — 35, 35 — 45, 45 — 55, etc.

How shall the step limits be indicated once they are fixed? There are four common methods of indicating step limits, but no matter which method is employed the important point is to remember just what the limits are, lest tabulation and statistical errors result. The conventional methods follow.

Where 6.0 is the lower step limit and the step interval is 1:

I Midpoint Method	Freq.	II Lower Limit Method	Freq.	III Both Limits Method	Freq.	IV Both Limits Method	Freq.
6.5	1	6.	1	6 — 7	1	6 — 6.999	1
7.5	3	7.	3	7 — 8	3	7 — 7.999	3
etc.	etc.	etc.	etc.	etc.	etc.	etc.	etc.

Where 5.5 is the lower step limit and the step interval is 2:

6.5	1	5.5	1	5.5 — 7.5	1	5.5 — 7.499	1
8.5	3	7.5	3	7.5 — 9.5	3	7.5 — 9.499	3
etc.	etc.	etc.	etc.	etc.	etc.	etc.	etc.

Method IV will cause fewest errors and is recommended for the beginner. Method II is most convenient and is recommended for use after the student has firmly fixed the habit of thinking of a score as spread over an interval. Method III will be used throughout this discussion, and the reader is specially warned to remember that 6 — 7 means 6 and up to *but not including* 7, i. e., 6 up to 6.999999999

and so on to infinity even though it will be written for convenience 6 — 7.

III. Order Distribution

Comparison and Construction.—The third mass measure is the simplest of all, and, like the frequency distribution, is the basis for certain types of subsequent statistical treatment. It is not so helpful, however, as either the frequency surface or the frequency distribution in the immediate study of the condition of a class.

If the scores in Table 30 are arranged in order of their size beginning with the smallest they appear as follows: 6, 7, 7, 8, 8, 8, 9, 9, 9, 9, 10, 10, 10, 10, 10, and so on up to 20. Such an arrangement would be an order distribution and would obviously be more intelligible than the arrangement in Table 30.

There is a fourth mass measure which is in some respects like the order distribution. This is the *rank distribution.* The construction and use of this measure will be considered later. Suffice it to say here that the construction of a rank distribution depends upon the preliminary construction of an order distribution. The order distribution shows not only what pupil made the highest score but also what was the actual numerical size of the score. The rank distribution neglects the numerical size of the score and merely states which pupil ranks highest, second, third and so on.

CHAPTER XV

STATISTICAL METHODS—POINT MEASURES

Kinds and Functions of Point Measures.—Mass measures may be vague and statistically cumbersome, but they possess the virtue of including every score in the class. It is the function of point measures to represent the condition of a class by a single number. The point chosen to represent the class depends upon the statistical method employed. The common methods are:

I. Mode.
II. Mean.[1]
III. Median or Midscore.
IV. Lower Quartile Point.
V. Upper Quartile Point.

I. Mode

What Is the Mode?—The *true* mode is too difficult in its computation for general use and will not be discussed further, but the *crude* mode is so simple that its calculation need not be specially illustrated. It is the simplest of all point measures, and gives a sort of look-and-see class score. *The crude mode is the most frequent score:* The mode of Fig. 26 is 13, of Fig. 28 is 15, of Fig. 29 is 7, of Fig. 30 is 11 and 18 since there are two modes, and of Table 32 is 7. Since 13 means 13 to 13.99, and similarly for the other scores, it is probably better to say the modes are at the midpoints, 13.5, 15.5, 7.5, etc.

[1] Throughout this book the term *average* is used in its generic sense to signify mode, mean, median, midscore, or any other measure of central tendency.

II. Mean

What Is the Mean?—A very common experience for many students is that they forget or get confused about all their ordinary arithmetic just as soon as they begin to study statistics. So it is well to state at the outset that the mean or arithmetic mean of statistics is the same good-old-school-day average. It is the *sum of the scores divided by the number of the scores.* Two differences will be noted. In the first place, when the mean of childhood was calculated from numbers like 12, 13, etc., 12 was 12 and wasn't 12-stretching-to-12.99, nor 12-wriggling-backward-to-11.5-and-squirming-forward-to-12.499, or at least we weren't conscious of 12 being all this. In the second place, statistics has developed certain shortcuts which are more or less novel, but which are very economical whenever the scores became numerous, even though they may appear more cumbersome for the following simple illustrations.

How to Compute the Mean.—Illustrative problems are worked out in Tables 32 and 33.

(a) *Scores Ungrouped* (*Table* 32)

(1) The scores are tabulated in an order distribution, though this is not necessary.

(2) The sum of the scores is 156 and the number of scores is 24.

(3) Then the mean $= \frac{156}{24} + .5 = 7.0$. This test is so scored that 2 means 2 — 2.999 and hence 2 is most truly represented by its mid-point 2.5 and similarly for all the other scores. This will make the mean .5 higher than $\frac{156}{24}$. This can be proved by adding up 2.5, 3.5, 4.5, etc., and by dividing this sum by 24.

TABLE 32

Number of Examples Done Correctly on Courtis Addition Test Series B

Scores Ungrouped

Pupil	Score
1	2
2	3
3	4
4	4
5	5
6	5
7	5
8	5
9	6
10	6
11	6
12	6
13	7
14	7
15	7
16	7
17	7
18	8
19	8
20	8
21	9
22	9
23	10
24	12

Sum $= 156$

$N = 24$

Mean $= \frac{156}{24} + .5$

$= 7.0$

Scores Grouped in Step Intervals of 1

Score	Frequency	Deviation from Guessed Mean	Freq. x Dev.
2—3	1	—5	— 5
3—4	1	—4	— 4
4—5	2	—3	— 6
5—6	4	—2	— 8
6—7	4	—1	— 4
			—27
7—8	5	0	
8—9	3	1	3
9—10	2	2	4
10—11	1	3	3
11—12	0	4	0
12—13	1	5	5
			15

$N = 24$

Guessed Mean $= 7.5$

15
-27
-12

Mean $= 7.5 + (-.5) \qquad \frac{-12}{24} = -.5$

$= 7.0$

(b) *Scores Grouped* (*Table* 32)

(1) The scores are retabulated in a frequency distribution. Previously most frequency distribu-

tions have been shown on the horizontal, but the vertical position is more convenient for statistical work.

(2) The sum of the frequencies or $N = 24$.

(3) Any step near the middle of the distribution is called the *guessed* mean. Guessed mean $= 7.5$. Any step may be chosen and the mean will come out the same.

(4) The scores are turned into deviation from the guessed mean. The step 6 — 6.99 is one point below (— 1) the guessed mean. Step 8 — 8.99 is one step above (+ 1) the guessed mean and so on.

(5) Each deviation is multiplied by its corresponding frequency. There is one deviation of — 5 and — 4. There are two deviations of — 3 making a total deviation of — 6 and so on.

(6) The sum of the minus deviations is — 27 and of the plus deviations + 15. The net sum is — 12, which when divided by N gives the correction — .5. This — .5 tells that the real mean is .5 below the guessed mean.

(7) The guessed mean is corrected to give the real mean. The commonest errors made in using the short method are: (a) failure to guess the mean at the mid-point of the step interval chosen; (b) failure to multiply the deviations by the frequencies; (c) a tendency to confuse the score column and the frequency column.

(a) *Scores Ungrouped* (*Table* 33)

(1) The scores are arranged in an order distribution.

(2) The sum of the scores is 1380 and N is 24.

(3) The mean $= \frac{1380}{24} = 57.5$. No correction is

TABLE 33

Quality of Penmanship as Judged with the Ayres Handwriting Scale

Scores Ungrouped		Scores Grouped in Step Intervals of 10			
Pupil	Score	Score	Frequency	Deviation from Guessed Mean	Freq. x Dev.
1	20	15—25	2	—30	—60
2	20				
3	40				
4	40	25—35	0	—20	— 0
5	40				
6	50				
7	50	35—45	3	—10	—30
8	50				
9	50				—90
10	50	45—55	6	0	
11	50				
12	60				
13	60	55—65	5	10	50
14	60				
15	60				
16	60	65—75	4	20	80
17	70				
18	70				
19	70	75—85	2	30	60
20	70				
21	80				
22	80	85—95	2	40	80
23	90				
24	90				270
Sum =	1380	N = 24			270
N =	24	Guessed Mean = 50			— 90
Mean =	$\frac{1380}{24} + 0$				180
		Mean = 50 + 7.5			$\frac{180}{24} = 7.5$
=	57.5	= 57.5			

added because the scores 20, 40, etc., are already at their mid-point. The values on the Ayres Handwriting Scale are 20, 30, 40, etc., with no intervening values.

As the scale is ordinarily used, a pupil's penmanship specimen which falls barely above 25 is called 30 and one barely below 35 is also called 30, hence 20 means 15 — 24.99 and 30 means 25 — 34.99. Thus the scores 20, 30, 40, etc., are, unlike the original scores in Table 32, already at the mid-point, and the mean needs no correction. If we imagine 20, 40, etc., to mean instead 20 — 29.99 and 40 — 49.99, the correction would be not .5 as in Table 32 but 5 since the mid-point of each score would be 5 higher than the beginning point.

(b) *Scores Grouped* (*Table* 33)

(1) The scores are retabulated in a frequency distribution. It will be observed that the step limits are made to fit the real meaning of the scores.

(2) $N = 24$. Guessed mean = 50, the mid-point of the 45 — 54.99 step.

(3) The deviation from the guessed mean of 35 — 44.99 is — 10. Note that the 25 — 34.99 step is inserted into the distribution. This must always be done.

(4) The remainder of the process is similar to that described for Table 32.

III. Lower Quartile, Median, and Upper Quartile

What Are the Median and Quartile Points?—These three point measures are treated together because of the similarity of their computation. The intimate relation of the three is shown by their definition. The *lower quartile* or Q_1 or *25 percentile is that point below which are 25% of the scores and above which are 75% of the scores. The median* or *50 percentile is that point below which are 50% of the scores and above which are 50% of the scores. The upper*

quartile or Q_3 or *75 percentile is that point which has 75% of the scores below it and 25% above it.*

Computation of Q_1 Median, and Q_3.—

TABLE 34

Number of Examples Done Correctly on Courtis Addition Test Series B

Scores Ungrouped

Pupil	Score	Computation
1	2	$\frac{N}{4} = 6\text{th}$
2	3	
3	4	$Q_1 = 5 + 2/4 \times 1$
4	4	
5	5	$Q_1 = 5.5$
6	5	
7	5	
8	5	$\frac{N}{2} = 12\text{th}$
9	6	
10	6	$\text{Median} = 7 + \frac{0}{5} \times 1$
11	6	
12	6	
13	7	$\text{Median} = 7$
14	7	
15	7	
16	7	
17	7	$3/4\,N = 18\text{th}$
18	8	
19	8	$Q_3 = 8 + 1/3 \times 1$
20	8	
21	9	$Q_3 = 8.33$
22	9	
23	10	
24	12	
$N = 24$		

Scores Grouped in Step Intervals of 1

Score	Frequency	Computation
2-3	1	$\frac{N}{4} = 6\text{th}$
3-4	1	$Q_1 = 5 + \frac{2}{4} \times 1$
4-5	2	$Q_1 = 5.5$
5-6	4	$\frac{N}{2} = 12\text{th}$
6-7	4	
7-8	5	$\text{Median} = 7 + \frac{0}{5} \times 1$
8-9	3	$\text{Median} = 7$
9-10	2	$3/4\,N = 18\text{th}$
10-11	1	$Q_3 = 8 + 1/3 \times 1$
11-12	0	$Q_3 = 8.33$
12-13	1	
$N = 24$		

(a) *Scores Ungrouped (Table 34)* — Q_1

(1) The scores are arranged in an order distribution.

(2) $N = 24$. How far down to count in order to locate Q_1 is shown by $\frac{N}{4}$. The 6th score then is the lower quartile point.

(3) The 6th score is 5. But since 5 means 5 — 5.99 just where between 5 and 5.99 is the Q_1? There are 4 scores spread between 5 — 5.99. Two of these 4 were used up in counting down 6 scores so the best guess is that Q_1 is 5 plus 2/4 of the distance 5 to 5.99. $Q_1 = 5.5$.

(b) *Scores Ungrouped — Median*

(1) $\frac{N}{2} = 12$. The 12th score down is the median.

(2) The 12th score uses up 0 of the five scores of 7, hence the median is 7 plus 0/5 of the distance from 7 to 7.99. Median = 7.0.

(c) *Scores Ungrouped — Q_3*

Three-fourths of $N = 18$. The 18th score is the Q_3. The 18th score uses up one of the three 8's. Therefore $Q_3 = 8 + 1/3 \times 1 = 8.33$.

(d) *Scores Grouped — Q_1*

(1) The scores are retabulated in a frequency distribution.

(2) $\frac{N}{4} = 6$. Hence the 6th score is the Q_1. The 6th score uses up frequencies, viz: 1 + 1 + 2 and 2 of the four 5 — 5.99's. Hence $Q_1 = 5 + \frac{2}{4} \times 1 = 5.5$. Thus the process is identical with that given for scores ungrouped.

(e) *Scores Grouped — Median*

The process is identical with that given for scores ungrouped.

(f) *Scores Grouped — Q_3*

The process is identical with that given for scores ungrouped.

The methods of calculating these three measures are practically identical whether the scores are grouped or ungrouped. When scores are grouped the counting down to

locate the three point measures is simplified, as is also the determination of the size of the correction. In locating the Q_1, for example, the 6th score uses up two of the four scores. The number 2 is placed as a numerator over the number 4, thus, $\frac{2}{4}$, and we have the correction as soon as the fraction is multiplied by the step interval. Multiplying by a step interval of 1 does not affect the correction, but it is well to establish the habit. Table 35 shows its necessity.

TABLE 35

Quality of Penmanship as Judged with the Ayres Handwriting Scale

Scores Ungrouped

Pupil	Score	Computation
1	20	$\frac{N}{4} = 6\text{th}$
2	20	
3	40	$Q_1 = 45 + 1/6 \times 10$
4	40	
5	40	$Q_1 = 46.67$
6	50	
7	50	
8	50	
9	50	
10	50	$\frac{N}{2} = 12\text{th}$
11	50	
12	60	Median $= 55 + 1/5 \times 10$
13	60	
14	60	Median $= 57$
15	60	
16	60	
17	70	
18	70	$3/4\,N = 18\text{th}$
19	70	
20	70	$Q_3 = 65 + 2/4 \times 10$
21	80	
22	80	$Q_3 = 70$
23	90	
24	90	
$N = 24$		

Scores Grouped in Step Intervals of 10

Score	Frequency	Computation
15-25	2	$\frac{N}{4} = 6\text{th}$
25-35	0	$Q_1 = 45 + 1/6 \times 10$; $Q_1 = 46.67$
35-45	3	
45-55	6	$\frac{N}{2} = 12\text{th}$
55-65	5	Median $= 55 + 1/5 \times 10$; Median $= 57$
65-75	4	
75-85	2	$3/4\,N = 18\text{th}$; $Q_3 = 65 + 2/4 \times 10$
85-95	2	$Q_3 = 70$
$N = 24$		

(a) *Scores Ungrouped (Table 35)* — Q_1

(1) The scores are arranged in an order distribution.

(2) $N = 24$. $\frac{N}{4} = 6$. The 6th score is Q_1. Counting down six scores, one of the six scores of 50 is used up. Hence the correction is 1/6 multiplied by the step interval 45 — 54.99. The correction, $1/6 \times 10$ is added to, not 50, but the beginning point of the score of 50, which is 45.

The calculation for Q_1 shows that the process is practically identical with that given for a performance test and this is true for the median and Q_3 as well. So beyond pointing out two slight differences, the illustration may be left to speak for itself. In the preceding problem the step interval was 1, in this it is 10, hence in this problem corrections are multiplied by 10. Again, in the first problem the score was expressed in its original form as the beginning point of the step, while in the second it was the mid-point. The frequency distributions in both problems make these two points clear.

Most of the difficulties which worry beginners in computing Q_1, Median, and Q_3 have now been considered. The computation of the median in Table 36 shows how to deal with a slightly different situation which is a frequent source of difficulty. There are some scales whose values are irregular. Such, for instance, is the Nassau County Extension of the Hillegas Composition Scale. Table 36 shows how to deal with such a situation. Note that the step limits are the half way points between scale values and that the size of the step interval fits the distances between scale values. In other words the step limits are made to harmonize with the way in which this product scale was scored. Also observe that even though N is an odd number the regular computation technique is unaffected. Also observe that

the upper limit of the last step interval cannot be given because the last value on the scale is 9.0.

TABLE 36

Scores According to the Nassau County Composition Scale.
Sample Scores: 0, 1.1, 1.9, 2.8, 3.8, 5.0, etc.

Quality	Frequency	Computation
.55 — 1.5	1	$\frac{N}{4} = \frac{19}{4} = 4.75$
1.5 — 2.35	2	$Q_1 = 2.35 + \frac{1.75}{3} \times .95$
2.35 — 3.3	3	$Q_1 = 2.9$
3.3 — 4.4	4	
4.4 — 5.5	5	$\frac{N}{2} = \frac{19}{2} = 9.5$
5.5 — 6.6	2	Median $= 3.3 + \frac{3.5}{4} \times 1.1$
6.6 — 7.6	0	Median $= 4.26$
7.6 — 8.5	1	$3/4\,N = 3/4$ of $19 = 14.25$
8.5 —	1	$Q_3 = 4.4 + \frac{4.25}{5} \times 1.1$
		$Q_3 = 5.34$
N =	19	

Another common source of difficulty is gaps in the scores, or rather, steps whose frequency is zero. Table 37 illustrates what to do when this difficulty is met. Note the procedure in computing Q_1. $\frac{N}{4} = 2$. Counting down the "Frequency" column 2 carries us through the step 2 — 4 but it doesn't carry us into the step 6 — 8. In other words Q_1 is between 4 and 6. When there is such a gap the best policy is to divide it equally between the step interval below and the step interval above. This has the effect of making the lower step interval 2 — 5 and the upper step interval 5 — 8. Consequently the correction is added to 5, the beginning point of the new step, and the correction is multi-

plied by 3, the size of the new step interval. Similarly in computing the median the beginning point of the step becomes 10, and the size of the step interval becomes 4. In computing Q_3, these numbers become 17 and 5 respectively.

TABLE 37

Scores According to the Monroe Standardized Fundamentals of Arithmetic Test Grouped in Step Intervals of 2

Score	Frequency	Computation
0—2	1	$\frac{N}{4} = \frac{8}{4} = 2$
2—4	1	$Q_1 = 5 + \frac{0}{2} \times 3$
4—6	0	$Q_1 = 5$
6—8	2	
8—10	0	$\frac{N}{2} = \frac{8}{2} = 4$
10—12	0	$\text{Median} = 10 + \frac{0}{2} \times 4$
12—14	2	$\text{Median} = 10$
14—16	0	$3/4\, N = 3/4 \text{ of } 8 = 6$
16—18	0	$Q_3 = 17 + \frac{0}{2} \times 5$
18—20	0	
20—22	2	$Q_3 = 17$
N =	8	

How to Compute the Midscore.—A small class of nine pupils made the following scores in an arithmetic test:

3, 9, 8, 7, 11, 12, 6, 3, 13.

The median of this series of scores, according to the computation method just described, is 8.5. The midscore on the other hand is 8.

The first step in computing a midscore is to arrange the scores in order of size. The above when so arranged are:

3, 3, 6, 7, 8, 9, 11, 12, 13.

The second step is to divide the number-of-scores-plus-one by two, thus:

$$(9 + 1) \div 2 = 5.$$

This tells us that the fifth score, counting from 3 toward 13 or from 13 toward 3, is the midscore. This gives a midscore of 8. When there is no midscore the midscore is usually considered as the mean of the two middlemost scores.

When to Use Each Average.—Use the mode when (a) quick computation is essential, or (b) the most frequent score is desired.

Use the mean when (a) every score should have an influence in determining the average which is exactly proportionate to the score's amount, or (b) when the lowest unreliability is sought, or (c) when subsequent correlation or other formulae or procedures require the mean.

Use the median when (a) quick computation is fairly important, or (b) a more popular average than the mean is desired, or (c) it is important that extreme or erroneous scores should not markedly influence the average, or (d) it is desired that certain scores exercise an influence in determining the average when all that is known concerning these scores is that they are above or below the average.

Use the midscore when (a) a very simple method of computation is required, or (b) the scores are discrete rather than continuous, i. e., they do not measure an infinitely continuous or divisible fact such as adding ability but a discrete indivisible fact such number of pupils and the like.

Usually the mean or median should be used.

CHAPTER XVI

STATISTICAL METHODS—VARIABILITY MEASURES

Need for Variability Measures.—The invariable fact in educational measurement is that pupils vary. Such variation, dispersion, or spread is a significant determiner of educational procedure. A measure of central tendency, while important, does not give a complete description of the condition of a class. Two classes when measured for spelling ability might show identical central tendencies with one varying from second-grade to eighth-grade ability and the other scarcely varying at all.

Nature of Variability Measures.—If we imagine the scores of a class to be tabulated in a frequency surface, the median or mean of this class is a mid-*point* and may be thought of as a point at or near the middle of the base line of the frequency surface. A variability measure on the other hand is not a point but a *distance,* just in the same way as an inch is a distance. The inch is a constant distance, but the variability measure is a variable distance, varying with the distribution for which it is calculated. A variability measure may be thought of as a certain distance along any part of the base line of a frequency surface. It is, however, most commonly thought of as being a certain distance just above or just below the central tendency.

Types of Variability Measures.—The mass measures —frequency surface, frequency distribution, and order distribution—already described give a far clearer picture of the variation within a class than do any other measures. If, however, it is desired to make use of variability in situations

where a single numerical value for it is required, one of the following conventional measures should be calculated.

I. *Total Range.* The total range distance includes 100% of the scores.

II. *Quartile Deviation (Q)* or *Semi-Interquartile Range.* A distance of Q above and a distance of Q below the central tendency includes roughly the middle 50% of the scores.

III. *Mean Deviation (Mn.D.* or *A.D.).* A distance of mean deviation above and below the central tendency includes roughly 57.5% of the scores.

IV. *Standard Deviation (S.D.)* or *Mean Square Deviation* or *Sigma* (σ). A distance of *S.D.* above and below the central tendency includes roughly 68% of the scores.

Neither the *Median Deviation (M.D.)* nor the *Probable Error (P. E.)* is included in the above list, nor will their calculation be illustrated at this point. The probable error will be considered later. For all practical purposes they may be considered equal to Q. In a normal distribution they are exactly so. It is better to reserve *P.E.* exclusively for use as an unreliability measure. The computation of the *M.D.* is, however, very simple. Just as a mean deviation is a mean of deviations regardless of signs, so an *M.D.* is a median of deviations without regard to signs.

Transmutation of One Variability Measure into Another.—The student will rarely need to compute more than one measure of variability, but if he does, and the distribution of scores is normal, all the other measures can be gotten from just one by means of the following. If the distribution is only approximately normal these relationships hold only approximately.

$$Q \text{ or } M.D. \text{ or } P.E. = .6745 \ S.D.$$

$$Mn.D. = .7979 \ S.D.$$

$$Q \text{ or } M.D. \text{ or } P.E. = .8453 \ Mn.D.$$

I. Total Range

Nature and Computation.—The total range is, as its name implies, the distance from the smallest score to the

largest score. It is computed by simply making the subtraction.

Like the mode in the simplicity of its computation, it is also like the mode in that it is valuable as an inspection measure only. This is so because it is peculiarly liable to large fluctuations, depending as it does upon but two scores.

II. Quartile Deviation

How to Compute *Q*.—Because of the identity of *Q* with *M.D.* or *P.E.* in a normal distribution; because of its satisfactory approximation to them in distributions which are not exactly normal; because $\pm Q$ above and below central tendency includes an easily understood middle 50% of scores, and because of the very great ease of its computation, the quartile deviation has become very popular.

The ease of its computation is shown by the following formula:

$$Q = \frac{Q_3 - Q_1}{2}$$

Thus it is half the distance from the lower quartile point to the upper quartile point. The scores in Table 34 yield a Q_3 of 8.33 and a Q_1 of 5.5. Then for this class

$$Q = \frac{8.33 - 5.5}{2} = 1.41 +$$

For Table 35

$$Q = \frac{70 - 46.67}{2} = 11.66 +$$

Thus the *Q* for each distribution is expressed in terms of its own distribution.

III. Mean Deviation

How to Compute the Mn.D.—The *Mn.D.* is *a mean of the deviations from any measure of central tendency, no account being taken of signs.* Table 38 and Table 39 illustrate the calculation of *Mn.D.* from the median. The

Mn.D., however, is more frequently computed from the mean.

TABLE 38

Number of Examples Done Correctly on Courtis Addition Test, Series B

Scores Ungrouped

Pupil	Score	Deviation from Median
1	2	−4.5
2	3	−3.5
3	4	−2.5
4	4	−2.5
5	5	−1.5
6	5	−1.5
7	5	−1.5
8	5	−1.5
9	6	− .5
10	6	− .5
11	6	− .5
12	6	− .5
13	7	.5
14	7	.5
15	7	.5
16	7	.5
17	7	.5
18	8	1.5
19	8	1.5
20	8	1.5
21	9	2.5
22	9	2.5
23	10	3.5
24	12	5.5

$N = 24$ Sum $= 42.0$
Median $= 7$
$Mn.D. = \frac{42}{24} = 1.75$

Scores Grouped in Step Intervals of 1

Score	Frequency	Deviation from Median	Frequency Times Deviation
2—3	1	−4.5	−4.5
3—4	1	−3.5	−3.5
4—5	2	−2.5	−5.0
5—6	4	−1.5	−6.0
6—7	4	− .5	−2.0
7—8	5	.5	2.5
8—9	3	1.5	4.5
9—10	2	2.5	5.0
10—11	1	3.5	3.5
11—12	0	4.5	0.0
12—13	1	5.5	5.5

$N = 24$ Sum $= 42.0$
Median $= 7$
$Mn.D. = \frac{42}{24} = 1.75$

(*a*) *Scores Ungrouped* (*Table* 38)

(1) The scores are arranged in an order distribution, though this is not necessary.

(2) $N = 24$, and the median according to previous calculation equals 7.

(3) The scores are expressed as deviations from the median. The first score of 2, meaning as it does 2 — 2.99, is best represented by its mid-point 2.5. The first score deviates from the median 4.5, the second by 3.5 and so on. The minus signs indicate deviations downward, but since the *Mn.D.* disregards signs they are not really needed.

(4) The sum of the deviations regardless of signs is 42.

(5) *Mn.D.* equals the sum of the deviations divided by *N*. $Mn.D. = \frac{42}{24} = 1.75.$

(*b*) *Scores Grouped*

(1) The scores are retabulated in a frequency distribution.

(2) The deviation of the first step, 2 — 2.99, is 4.5, of the second step, 3.5 and so on. These deviations are not *step* deviations but actual deviations.

(3) The deviations are multiplied by their corresponding frequencies. There is one deviation of 4.5, one of 3.5, two of 2.5 which means a total deviation of 5.0, etc.

(4) The sum of the deviations regardless of signs is 42.

(5) $Mn.D. = \frac{42}{24} = 1.75.$

The student will find Table 39 self-explanatory, for there is no difference from the method described above, except that the ungrouped scores are already at their mid-point.

Mn.D. vs. Q.—It was pointed out a few pages back that $\pm$ *Q* includes 50% of the scores and $\pm$ *Mn.D.* includes 57.5% of the scores. If this is so, the *Mn.D.* should be larger than *Q* for the same distribution. The illustrative problems reveal just such a relationship.

TABLE 39

Quality of Penmanship as Judged by the Ayres Handwriting Scale

Scores Ungrouped			Scores Grouped in Step Intervals of 10			
Pupil	Score	Deviation from Median	Score	Frequency	Deviation from Median	Frequency Times Deviation
1	20	−37	15—24.9	2	−37	−74
2	20	−37				
3	40	−17	25—34.9	0	−27	00
4	40	−17				
5	40	−17				
6	50	− 7	35—44.9	3	−17	−51
7	50	− 7				
8	50	− 7				
9	50	− 7	45—54.9	6	− 7	−42
10	50	− 7				
11	50	− 7				
12	60	3	55—64.9	5	3	15
13	60	3				
14	60	3				
15	60	3	65—74.9	4	13	52
16	60	3				
17	70	13				
18	70	13	75—84.9	2	23	46
19	70	13				
20	70	13				
21	80	23	85—94.9	2	33	66
22	80	23				
23	90	33				
24	90	33				

Scores Ungrouped: $N = 24$ Sum $= 346$; Median $= 57$; $Mn.D. = \frac{346}{24} = 14.416+$

Scores Grouped: $N = 24$ Sum $= 346$; Median $= 57$; $Mn.D. = \frac{346}{24} = 14.416+$

Problem I. $Q = 1.41$. $Mn.D. = 1.75$
Problem II. $Q = 11.16$. $Mn.D. = 14.42$

IV. Standard Deviation

How to Compute the *S.D.*—Like the *Mn.D.*, the *S.D.* may be computed from any average or measure of central

tendency whether mode, median, or mean. Tables 40 and 41 illustrate its calculation from the mean.

TABLE 40

Number of Examples Done Correctly on Courtis Addition Test, Series B

Scores Ungrouped

Pupil	Score	Deviation from Guessed Mean	Deviation Squared
1	2	—5	25
2	3	—4	16
3	4	—3	9
4	4	—3	9
5	5	—2	4
6	5	—2	4
7	5	—2	4
8	5	—2	4
9	6	—1	1
10	6	—1	1
11	6	—1	1
12	6	—1	1
13	7	0	0
14	7	0	0
15	7	0	0
16	7	0	0
17	7	0	0
18	8	1	1
19	8	1	1
20	8	1	1
21	9	2	4
22	9	2	4
23	10	3	9
24	12	5	25

N = 24 Sum = 124
Mean = 7.0
Guessed Mean = 7.5

$$S.D. = \sqrt{\frac{124}{24} - (7.5 - 7.0)^2} = 2.217 +$$

Scores Grouped in Step Intervals of 1

Score	Frequency	Deviation from Guessed Mean	Frequency Times Deviation^2
2— 3	1	—5	25
3— 4	1	—4	16
4— 5	2	—3	18
5— 6	4	—2	16
6— 7	4	—1	4
7— 8	5	0	0
8— 9	3	1	3
9—10	2	2	8
10—11	1	3	9
11—12	0	4	0
12—13	1	5	25

N = 24 Sum = 124
Mean = 7.0
Guessed Mean = 7.5

$$S.D. = \sqrt{\frac{124}{24} - (7.5 - 7.0)^2} = 2.217 +$$

(*a*) *Scores Ungrouped* (*Table* 40)

(1) The scores are arranged in an order distribution, though this is not necessary.

(2) $N = 24$, and according to previous calculation the mean = 7.0. In order to avoid decimals in the deviations a guessed mean of 7.5 is used instead of the mean 7.0. A guessed mean of 9.5 or 2.5 would serve just as well.

(3) Each score is expressed as a deviation from the guessed mean.
(4) Each deviation is squared.
(5) The sum of the squared deviations is 124.
(6) The *S.D.* equals the square root of the sum of the squared deviations divided by *N*, minus the correction squared. The correction is the difference between the mean and the guessed mean, in this case .5.

$$S.D. = \sqrt{\frac{124}{24} - (7.5 - 7.0)^2}$$

(*b*) *Scores Grouped*
(1) The scores are retabulated in a frequency distribution.
(2) $N = 24$ and the mean $= 7$.
(3) The mid-point of any step near the middle of the distribution is taken as the point of reference. This guessed mean is always guessed at the *mid-point* of the step chosen. Guessed mean $= 7.5$.
(4) Each step is expressed as an actual deviation from the guessed mean.
(5) Each deviation is squared and then multiplied by its corresponding frequency. Beginning at the top, this gives results, viz: $(5)^2 \times 1 = 25$, $(4)^2 \times 1 = 16$, $(3)^2 \times 2 = 18$, etc.
(6) The sum of the squared deviations is 124.
(7) The $S.D. = \sqrt{\frac{124}{24} - (\text{correction})^2}$. The correction is the difference between the guessed mean and the actual mean, in this case .5. In case there is no difference the correction is zero. The advantage of computing deviations from the guessed mean and then correcting, instead of from the actual mean, is because the deviations can always be kept in whole numbers.

After the explanation given above the sample problems worked out below need not be described. Note that the answer is the same whether 50 or 60 is taken as the guessed mean.

TABLE 41

Quality of Penmanship as Judged with the Ayres Handwriting Scale

Scores Ungrouped

Pupil	Score	Deviation from Guessed Mean	Deviation Squared
1	20	—30	900
2	20	—30	900
3	40	—10	100
4	40	—10	100
5	40	—10	100
6	50	0	0
7	50	0	0
8	50	0	0
9	50	0	0
10	50	0	0
11	50	0	0
12	60	10	100
13	60	10	100
14	60	10	100
15	60	10	100
16	60	10	100
17	70	20	400
18	70	20	400
19	70	20	400
20	70	20	400
21	80	30	900
22	80	30	900
23	90	40	1600
24	90	40	1600

$N = 24$ Sum $= 9200$
Mean $= 57.5$
Guessed Mean $= 50$

$$S.D. = \sqrt{\frac{9200}{24} - (57.5 - 50)^2} = 18.085$$

Scores Grouped in Step Intervals of 10

Score	Frequency	Deviation from Guessed Mean	Frequency Times Deviation²
15—25	2	—40	3200
25—35	0	—30	00
35—45	3	—20	1200
45—55	6	—10	600
55—65	5	0	00
65—75	4	10	400
75—85	2	20	800
85—95	2	30	1800

$N = 24$ Sum $= 8000$
Mean $= 57.5$
Guessed Mean $= 60$

$$S.D. = \sqrt{\frac{8000}{24} - (60 - 57.5)^2} = 18.085$$

The *S.D.* from a Median.—Because the *S.D.* may be computed from any average there is a popular supposition that the *S.D.* from a mean and a median are computed in exactly the same way. But deviations cannot be computed from a guessed median as they can be from a guessed mean, and corrected for in the formula

$$S.\,D. = \sqrt{\frac{\text{Sum of (dev.)}^2}{N} - (\text{correction})^2}.$$

This formula holds for the mean only. A frequency distribution can be used but all deviations *must be from the actual median,* in which case the formula is identical with the above with the correction omitted.

***S.D. vs. Mn.D.* or *Q*.**—The per cents of the scores included by $\pm$ *Q*, $\pm$ *Mn.D.*, and $\pm$ *S.D.* are respectively, for a normal distribution, 50%, 57.5%, and 68%. Consequently there should be an increase in the sizes of the respective variability measures. The facts are as follows:

Problem I.	$Q =$ 1.41.	*Mn. D.* = 1.75	*S. D.* = 2.22
Problem II.	$Q =$ 11.66.	*Mn. D.* = 14.42	*S.D.* = 18.085

The transmutation table given a few pages back shows that when the frequency distribution is normal $Q = .6745$ *S.D.*, *Mn.D.* = .7979 *S.D.*, and $Q = .8453$ *Mn.D.* Had *S.D.* only been computed and transmuted into *Mn.D.* and *Q*, in Problem I, *Mn.D.* would have been $2.22 \times .7979$, or 1.77, instead of 1.75, and *Q* would have been $2.22 \times .6745$, or 1.497, instead of 1.41. The transmutation formulæ do not yield exactly the same results as calculation because we are not dealing with perfectly normal frequency distributions.

When to Use Each Variability Measure.—Use *Total Range* (a) for inspection purposes only, or (b) as a supplement to other variability measures.

Use *Q* (a) when an easily and quickly computed measure which is reasonably satisfactory is desired, or (b) when Q_1 and Q_3 will be important supplementary information.

Use *S.D.* (a) when it is desired to allow extreme scores to markedly influence the variability measure, or (b) when low unreliability is desired, or (c) when subsequent correlation or reliability formulæ require the *S.D.*

The *Q* and *S.D.* will suffice for practically every situation. There is little justification for continuing the use of either *Mn.D.* or *M.D.*, and *P.E.* should be definitely reserved as a measure of reliability rather than variability.

CHAPTER XVII

STATISTICAL METHODS—RELATIONSHIP AND RELIABILITY MEASURES

I. Relationship Measures

What Is Correlation?—The idea of correlation is so familiar that it is found in literary masterpieces and in the fables of the street. This is especially the case with inverse or negative correlation. "For every grain of wit there is a grain of folly." "The vulnerable heel of Achilles." "The leaf spot of Siegfried." "Beauty *vs.* Brains." "Eye-minded *vs.* ear-minded." "Idea thinkers *vs.* thing thinkers."

Thus correlation is a method for determining *the correspondence and proportionality between two series of scores or measures for the same pupils,* or the same schools, or the same cities, or any other entity. When the correspondence is perfect and positive the coefficient of correlation (r) is $+ 1.0$, when it is perfect, but negative, r is $- 1.0$. Correlation is *positive* when one series of scores tends to increase as the other increases, and *negative* when one tends to increase as the other decreases. A coefficient of correlation may be any size from $+ 1.0$ through 0 to $- 1.0$.

Pupil	Test I Score	Test II Score	Test I Score	Test III Score	Test I Score	Test IV Score	Test I Score	Test V Score
A	2	6	2	12	2	6	2	12
B	3	8	3	10	3	10	3	8
C	4	10	4	8	4	8	4	10
D	5	12	5	6	5	12	5	6
	$r = + 1.0$		$r = - 1.0$		$r = + .8$		$r = - .8$	

Some Uses of Correlation.—Here are some of the questions which education often asks and correlation can

answer: How reliable is this mental or educational test? Does increasing its length or repeating it increase its reliability? Do these two tests measure the same aspect of reading ability, as they claim? Which one of a group of tests is most representative of all of them? Is there any justification for the popular assumption that pupils who are best in English tend to be poor in mathematics? Do those who work most rapidly in arithmetic tend to work most accurately? How reliable is a teacher's examination in history? How close is the agreement between a test and a teacher's judgment? How close is the agreement between school marks and success in life? These and hundreds of other such questions involving a relationship between two series of measures can be answered by correlation.

Here are a few statements that correlation cannot make: When correlation is .8, 80% of the pupils show perfect correspondence. When correlation is positive but less than perfect a larger score in one series *always* accompanies a larger score in the other series. When there is a high correlation between two series of facts one has caused the other, or correlation implies causal relation.

How to Compute Correlation by the Standard Method.—There are several excellent methods for computing a coefficient of correlation. The product-moment method is the one most commonly used and generally approved. The product-moment formula for calculating a coefficient of correlation is,

$$r = \frac{\Sigma xy}{N \sigma_x \sigma_y}$$

which may be stated in this form,

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 . \Sigma y^2}}$$

This formula is made clear by the simple problem in Table 42 with its explanation.

TABLE 42

Correlation Between the Scores of a Class on Courtis Addition Test, Series B, and Ayres Handwriting Scale (Standard Method)

Pupil	Score I	Score II	Dev. from Mean I x	Dev. from Mean II y	x^2	y^2	xy
A	2	50	—5	— 7.5	25	56.25	37.5
B	3	50	—4	— 7.5	16	56.25	30.0
C	4	50	—3	— 7.5	9	56.25	22.5
D	4	80	—3	22.5	9	506.25	— 67.5
E	5	20	—2	—37.5	4	1406.25	75.0
F	5	60	—2	2.5	4	6.25	— 5.0
G	5	40	—2	—17.5	4	306.25	34.0
H	5	50	—2	— 7.5	4	56.25	15.0
I	6	70	—1	12.5	1	156.25	— 12.5
J	6	40	—1	—17.5	1	306.25	17.5
K	6	70	—1	12.5	1	156.25	— 12.5
L	6	50	—1	— 7.5	1	56.25	7.5
M	7	50	0	— 7.5	0	56.25	0.0
N	7	70	0	12.5	0	156.25	0.0
O	7	40	0	—17.5	0	306.25	0.0
P	7	70	0	12.5	0	156.25	0.0
Q	7	60	0	2.5	0	6.25	0.0
R	8	20	1	—37.5	1	1406.25	— 37.5
S	8	60	1	2.5	1	6.25	2.5
T	8	90	1	32.5	1	1056.25	32.5
U	9	80	2	22.5	4	506.25	45.0
V	9	60	2	2.5	4	6.25	5.0
W	10	90	3	32.5	9	1056.25	97.5
X	12	60	5	2.5	25	6.25	12.5
Mean	7.	57.5	Sum or Σ		124	7850.00	434.0
							—135.0
							299.0

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} = \frac{299}{\sqrt{(124)(7850)}} = \frac{299}{986.61} = .303$$

(1) The scores for tests I and II are paired according to the pupils making them.

(2) By previous calculation the mean of test I is 7.0 and of test II is 57.5. Theoretically only the mean may be used in product-moment correlation, but in practice the median is sometimes used.

(3) The scores from tests I and II are turned into deviations from their

own mean. The first column of deviations is called x and the second y.

(4) Each x and each y is squared. The square of -5 is 25, of -7.5 is 56.25 and so on down.

(5) Each x is multiplied by its corresponding y. The product of -5 and -7.5 is $+37.5$ and so on down.

(6) The sum of the x^2's and the y^2's is computed.

$$\Sigma x^2 = 124$$
$$\Sigma y^2 = 7850$$

(7) The sum of the plus xy's $= 434$, and of the minus xy's $= 135$. The algebraic sum of the two is determined. $\Sigma xy = 299$.

(8) Substituting these values in the formula and solving, $r = .303$.

How to Compute Correlation by the Method of Ranks.—In a serious correlation study the student should use the standard formula, but when time is an important consideration and refinements are not essential, he may use the Spearman "Footrule" formula

$$R = 1 - \frac{6 \Sigma G}{N^2 - 1}$$

and transmute R into r by means of Table 44. Pearson has shown that the true correlation is not approximated by Spearman's R until it is transmuted into r by Table 44, which is constructed from Pearson's formula,

$$r = 2 \cos \frac{\pi}{3} (1 - R) - 1$$

The problem in Table 42 is recalculated by the Spearman method of ranks in Table 43.

(1) The scores of test I are each given a relative rank. The score 2 is ranked first or 1, score 3 is ranked 2, score 4 is ranked 3.5 because the two scores of 4 occupy ranks 3 and 4 whose average is 3.5; score 5 is ranked 6.5 since the four scores of 5 occupy ranks 5, 6, 7, and 8, whose average is 6.5 and so on for the other scores. The scores in test II are also ranked in order beginning with the smallest score. There are two scores of 20 occupying ranks 1 and 2 so each 20 is ranked 1.5. The three scores of 40 occupy ranks 3, 4, and 5, whose average is 4 and so on for the remaining scores of test II. The largest score in tests I and II might have received the rank of 1 instead of the smallest, and in fact this is often done. Either method of ranking is correct provided the method is uniform for both tests.

(2) Compute the gains in rank of test II over test I. Thus we have $8.5 - 1 = 7.5$; $8.5 - 2 = 6.5$, etc.

(3) The sum of the gains in rank is computed. $\Sigma G = 74.5$.

(4) Substituting values in the formula and solving $R = .224$. Transmuting R by Table 44, $r = .37$.

TABLE 43

Correlation Between the Scores of a Class on Courtis Addition Test Series B and Ayres Handwriting Scale (Rank Method)

Pupil	Score I	Score II	Rank I	Rank II	Gain in Rank G
A	2	50	1	8.5	7.5
B	3	50	2	8.5	6.5
C	4	50	3.5	8.5	5.0
D	4	80	3.5	21.5	18.0
E	5	20	6.5	1.5	
F	5	60	6.5	14	7.5
G	5	40	6.5	4	
H	5	50	6.5	8.5	2.0
I	6	70	10.5	18.5	8.0
J	6	40	10.5	4	
K	6	70	10.5	18.5	8.0
L	6	50	10.5	8.5	
M	7	50	15	8.5	
N	7	70	15	18.5	3.5
O	7	40	15	4	
P	7	70	15	18.5	3.5
Q	7	60	15	14	
R	8	20	19	1.5	
S	8	60	19	14	
T	8	90	19	23.5	4.5
U	9	80	21.5	21.5	
V	9	60	21.5	14	
W	10	90	23	23.5	.5
X	12	60	24	14	

$N \quad 24 \qquad \Sigma G = 74.5$

$$R = 1 - \frac{6\,\Sigma G}{N^2 - 1} = 1 - \frac{6\,(74.5)}{(24)^2 - 1} = .224$$

By Table 44, r = .37

How to Interpret a Correlation Coefficient.—Is an *r* of .30 or .37, according to the formula used, "high" or "low"? With *r*'s as with intelligence, or wealth, or beauty, the customary criterion is that of relativity. There seems to be a sort of rough agreement among workers in this field that when *r* is

TABLE 44

Transmutation of R into r according to

$$r = 2\cos\frac{\pi}{3}(1 - R) - 1. \quad R = 1 - \frac{6\,\Sigma\,G}{N^2 - 1}$$

R	r	R	r	R	r	R	r
.00	.000	.26	.429	.51	.742	.76	.937
.01	.018	.27	.444	.52	.753	.77	.942
.02	.036	.28	.458	.53	.763	.78	.947
.03	.054	.29	.472	.54	.772	.79	.952
.04	.071	.30	.486	.55	.782	.80	.956
.05	.089	.31	.500	.56	.791	.81	.961
.06	.107	.32	.514	.57	.801	.82	.965
.07	.124	.33	.528	.58	.810	.83	.968
.08	.141	.34	.541	.59	.818	.84	.972
.09	.158	.35	.554	.60	.827	.85	.975
.10	.176	.36	.567	.61	.836	.86	.979
.11	.192	.37	.580	.62	.844	.87	.981
.12	.209	.38	.593	.63	.852	.88	.984
.13	.226	.39	.606	.64	.860	.89	.987
.14	.242	.40	.618	.65	.867	.90	.989
.15	.259	.41	.630	.66	.875	.91	.991
.16	.275	.42	.642	.67	.882	.92	.993
.17	.291	.43	.654	.68	.889	.93	.995
.18	.307	.44	.666	.69	.896	.94	.996
.19	.323	.45	.677	.70	.902	.95	.997
.20	.338	.46	.689	.71	.908	.96	.998
.21	.354	.47	.700	.72	.915	.97	.999
.22	.369	.48	.711	.73	.921	.98	.9996
.23	.384	.49	.721	.74	.926	.99	.9999
.24	.399	.50	.732	.75	.932	1.00	1.0000
.25	.414						

0 to $\pm$ .4 correlation is low, or
$\pm$ .4 to $\pm$ .7 correlation is substantial, or
$\pm$ .7 to $\pm$ 1.0 correlation is high.

There is, however, a more satisfactory way to interpret coefficients of correlation. When we have perfect correlation between two traits it is possible to predict accurately an individual's position in one of these traits from a knowledge of his position in the other. As the coefficient of correlation goes toward zero such predictions become more and more uncertain. When the coefficient is exactly zero a pre-

diction has no more accuracy than a sheer guess or a purely chance estimate. Kelley has worked out the data of Table 45. According to this table, when $r = 0$ the error of prediction is 1.00, where 1.0 is defined as a sheer guess. When $r = .1$ the error has been reduced to .995. The coefficient of correlation must be about .85 before the error is half way between a guess and perfect prediction. Slight increases in the size of the coefficient above this point cause a rapid decrease in the error of prediction.

TABLE 45

Shows Decreases in the Error of Prediction from 1.00 toward Zero with Increases in *r* from Zero toward 1.0, Where an Error of 1.00 Is a Sheer Guess and an *r* of 1.00 Is Perfect Correlation

r	Error
.00	1.000
.10	.995
.20	.9798
.30	.9539
.40	.9165
.50	.8660
.60	.8000
.70	.7141
.80	.6000
.85	.5268
.90	.4359
.95	.3122
.97	.2431
.99	.1411

Equal in importance to the highness or lowness of an *r* is its reliability. As will be seen presently this is determined by the size of the *r* and the number of pupils. An *r* of .30 from only 24 pupils is relatively very low, has an error of prediction of .954, and is besides very unreliable.

When May Correlation Be Applied?—Both of the formulæ given assume a *rectilinear* or *straight line* relationship. Fig. 33 shows how to plot the scores of Table 43 to show whether their relationship is rectilinear or curvilinear.

In so far as there is any drift of the points at all, the drift is forward and upward roughly in a straight line direction. The great dispersion of the points indicates low correlation. Fig. 34 shows rectilinear relationship coupled with

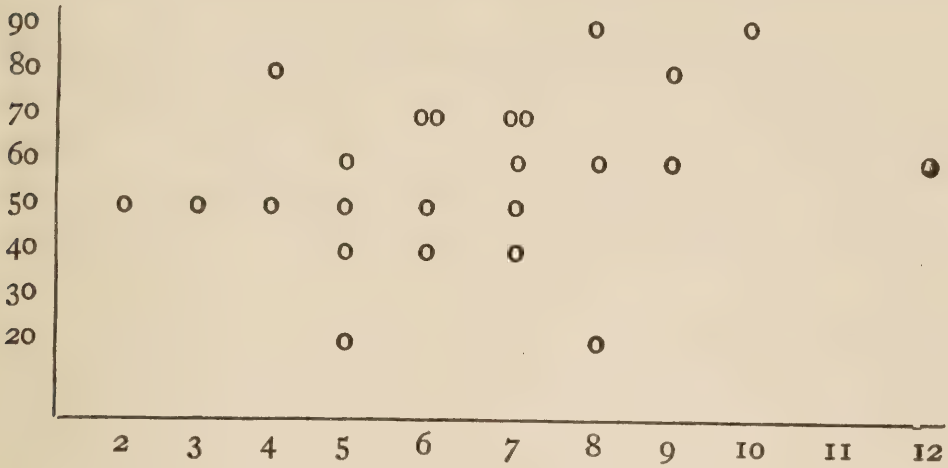


FIG. 33. Shows Rectilinear Relationship with an r of .303. (Data from Table 43.)

higher correlation. Fig. 35 shows three kinds of problems plotted on one pair of axes, (A) perfect positive correlation, (B) perfect negative correlation, (C) curvilinear relationship which cannot be treated by methods given in this book.

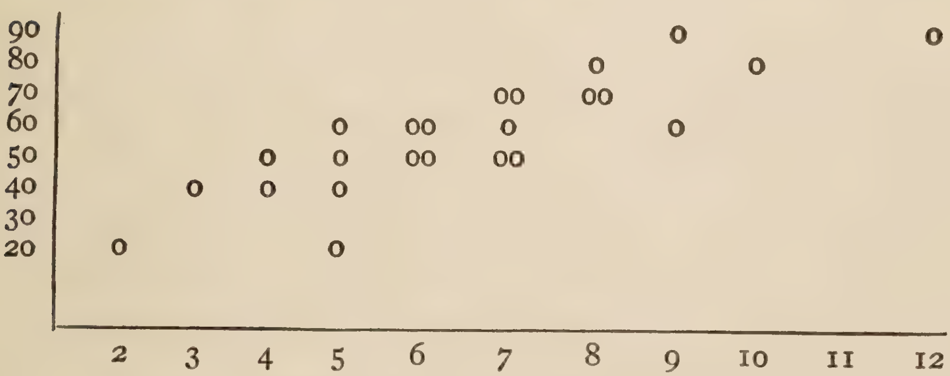


FIG. 34. Shows Rectilinear Relationship with an r of .8

Self-Correlation Coefficients.—Self-correlation is the correlation between two duplicate tests given to the same pupils. Its chief function is to show whether one test is a sufficently accurate measure of each pupil. Reliability is

one criterion for evaluating a test. Self-correlation is one statistical technique whereby a test's reliability may be determined. If the self-correlation between two duplicate tests is 1.0, then one test is an absolutely accurate measure of each pupil in the trait which the test measures. This ideal is of course never attained.

How high should self-correlation be? No absolute standard can be given that will fit every situation. Where test results are used to commit children to institutions or to exclude them from important social or educational opportunities and the like, or where results are to be used for close

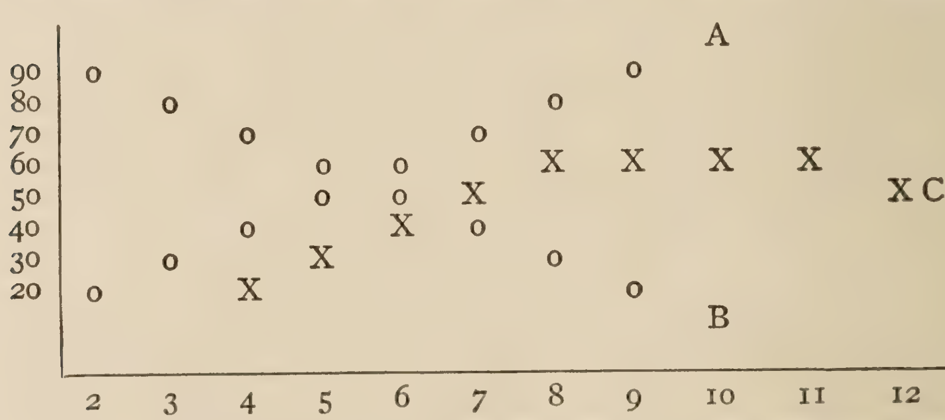


FIG. 35. A—Rectilinear Relationship with an r of + 1.0. B—Rectilinear Relationship with an r of — 1.0. C—Curvilinear Relationship.

theoretical reasoning self-correlation should certainly be above .9. But such a criterion is too drastic for most practical purposes, since the range of self-correlation for most standard tests is about .5 to about .9, while the range for typical teachers' examinations is much lower. A criterion of .9 or above would disqualify most educational tests and forbid as a public nuisance a professor's examination. Clara Chassell has found that the self-correlation of the marks of college professors on students who were rated through four full years is only .80! If the coefficient is not satisfactorily high it is evidence that one of two things needs to be done: (a) The test must be lengthened. How much it must be lengthened can be determined by computing the new correlation between the lengthened test and a duplicate of

it. (b) If the test is not lengthened or not lengthened enough it must be repeated. How many times to repeat can be determined empirically by giving a test and its duplicate twice each and correlating the two series of averages, and if that is not enough, by giving each test three times and correlating averages, etc.

But this empirical process is very expensive in time, since twice as many tests as are needed must be given before it can be determined just how many are needed. The use of Spearman's prophecy formula will save half of this time.

$$r_x = \frac{N\, r_1}{1 + (N - 1)\, r_1}$$

If the self-correlation of one test with a duplicate (r_1) is .8, and the information sought is how many times (N) the test must be given to yield a desired coefficient (r_x) of .9, substitute as follows and solve for N:

$$.9 = \frac{N\,(.8)}{1 + (N - 1)\,.8}$$

$$N = 2.25 \text{ times}$$

If the information sought is the r_x which would result from giving the same test or similar tests four times, substitute as follows and solve for r_x:

$$r_x = \frac{4\,(.8)}{1 + (4 - 1)\,.8} = .941$$

Suppose that r_1 or .8 were the self-correlation between the average of two duplicate tests and the average of two other similar tests. In that case the N required to yield a self-correlation of .9 would be 2.25×2 or 4.5. The second formula would be interpreted as follows: 4 pairs of tests or 8 duplicate tests in all will yield an r_x of .941.

Other Relationship Measures.—Correlation is not the only method for computing relationship. It is probable that the beginner will do better to compute relationship as follows: (1) Express each of the two series of scores as a deviation from its own average. (2) Divide these deviations in each series by the *S.D.* of that series in order to equate

variability. (3) Find the difference between the two equated deviations for each pupil. In doing this have regard for signs. In case there is a perfect relationship all these differences will be zero. Any difference larger than zero shows the amount of displacement in terms of *S.D.* (4) Make a frequency distribution of the differences. (5) Compute the mean or median to determine the average amount of *S.D.* displacement.

II. Reliability Measures

What Are Unreliability Measures?—Imagine that in a certain city there are 1000 pupils in the sixth grade. We wish to know the mean on an arithmetic test for the entire 1000, but there is only time to measure 100 of them. If 100 pupils are selected as nearly as possible at random from the 1000, and these 100 are measured, and the mean and *S.D.* of their scores is computed, it is possible by means of an unreliability formula to discover the limits within which the *true* mean for the entire 1000 will fall. Similarly, by use of the appropriate formula, it is possible to determine the limits within which the true median, or the true *Q* or the true *S.D.* for the entire 1000 pupils will fall.

But to better understand the unreliability formulæ, imagine again that the 1000 pupils are measured in ten groups of 100 each selected as nearly at random as possible. This will yield ten means, no one of which will probably agree exactly with any other or with the mean of all ten which is the *true* mean. Just as it is possible to compute (a) the mean and (b) *S.D.* for 100 pupil scores, so it is possible to compute (c) the mean of the ten means, and (d) the *S.D.* of the ten means. These four measures are called respectively, (a) *obtained mean,* (b) *S.D. distribution,* (c) the *true mean,* and (d) *S.D. mean. S.D. mean* is a measure of the unreliability of any one of the ten means for it is a measure of the variability among the ten means and hence is an index of *each one's* most probable divergence from the

true mean. *S.D. mean,* then, measures the unreliability not of the mean of the ten means for, since it is the true mean, it has no unreliability, but of the unreliability of any one of the ten means.

To illustrate, suppose that actual measurement of the 1000 pupils in groups of 100 showed the following:

	Mean Score	*S. D. Distribution*
For the first 100 pupils	25	9
" " second " "	23	10
" " third " "	24	12
" " fourth " "	27	14
" " fifth " "	25	10
" " sixth " "	26	11
" " seventh " "	24	13
" " eighth " "	26	12
" " ninth " "	25	11
" " tenth " "	25	8

Then

True mean = 25. True *S.D. distribution* = 11.
S.D. mean = 1.1 *S.D. S.D.* = 1.7

Just as *S.D. mean,* i. e., the reliability of any one mean from 100 pupils, was computed from the ten means, and just as *S.D. S.D.*, i. e., the reliability of any one *S.D. distribution* was computed from the ten *S.D. distribution,* so it would be possible to compute in similar fashion a reliability measure for a median (*S.D. median*), or for a *Q* (*S.D. Q*), or for an *r* (*S.D. r*), and so on.

The function of reliability formulæ is to prophecy from just one sampling what these *S.D.'s* would be if an individual were to go through the labor of testing all the possible samplings (in the above case 10). Thus the reliability formula for the mean is

$$S.D.\ mean = \frac{S.D.\ distribution}{\sqrt{N}}$$

Suppose, instead of testing ten groups of 100 pupils and actually determining empirically the *S.D. mean,* that just

the first 100 had been tested and the *S.D. distribution* computed, and the *S.D. mean* determined through the reliability formula. This would have been a great saving of time and would have given an *S.D. mean* only two-tenths in error as compared with the true *S.D.* mean of 1.1. Thus:

$$S.D.\ mean = \frac{9}{\sqrt{100}} = .9$$

Suppose, for illustration, we enquire about the unreliability of some of the measures already calculated. One of the problems carried throughout recent chapters is summarized in Table 46.

TABLE 46

Frequency Distribution of the Number of Examples Done Correctly on Courtis Addition Test Series B, together with Certain Statistical Measures Which Have Previously Been Calculated

Score	Frequency	Statistical Results
2—3	1	Mean = 7.0
3—4	1	Median = 7.0
4—5	2	
5—6	4	Q = 1.4
6—7	4	
7—8	5	$S.\ D.$ = 2.22
8—9	3	
9—10	2	r = .303 (with Ayres Handwriting)
10—11	1	
11—12	0	
12—13	1	
N	24	

Unreliability of the Mean in Table 46.—The unreliability of mean 7.0 is shown by the following formula:

$$S.D.\ mean = \frac{S.D.\ distribution}{\sqrt{N}}$$

$$S.D.\ mean = \frac{2.22}{\sqrt{24}} = .45$$

To say that the unreliability of the mean or *S.D. mean* is .45 is not particularly illuminating to most people. But unreliability can be stated more intelligently. It is customary to call *practical certainty* ± 3 *S.D. measure.* If we are concerned with the mean, practical certainty is ± 3 S.D. *mean,* if our concern is with the median, practical certainty is ± 3 *S.D. median,* if with the *S.D.* it is ± 3 *S.D.* S.D. and so on.

We can be practically certain, then, that the true mean is between 7.0 — 3 (.45) and 7.0 + 3 (.45), or, as it will be written hereafter, 7.0 ± 3 (.45), i. e., we can be practically certain that the true mean is somewhere between 5.65 and 8.35.

But the true mean of *what* falls between 5.65 and 8.35? It is not the true mean of the 24 pupils, because by actual computation we know that their true mean is 7.0. It is the true mean of any larger group of which the 24 pupils are an attempted random sampling. If these 24 pupils were a *perfectly* random sampling of some larger group the mean 7.0 would be the exact mean for the larger group. But we can scarcely hope to make a perfect sampling. Let us assume that the 24 pupils are, so far as can be made, a random sampling, or chance selection from all fifth-grade pupils in New York City. According to the data we can feel assured that the true mean for all New York City fifth-grade pupils is somewhere between 7.0 ± 3 (.45). There is always a possibility that the true mean is below 5.65 or above 8.35, but the chance of this being so is exceedingly small. The chances are in fact only 3 in 10,000. We cannot be practically certain, but the chances are very great that the true mean is between 7.0 ± 2 (.45). The chances are substantial that the true mean is between 7.0 ± 1 (.45). Just what the chances are will be shown later. The above treatment applies to *S.D. measure,* i. e., to any measure of unreliability. The formulæ for the most important of these follow.

Unreliability of the Median in Table 46.—

$$S.D.\ median = \frac{1\frac{1}{4}\ S.D.\ distribution}{\sqrt{N}}$$

$$S.D.\ median = \frac{1\frac{1}{4} \times 2.22}{\sqrt{24}} = .57$$

We can be practically certain that the true median of the group of which the 24 pupils are a random sampling is between 7.0 $\pm$ 3 (.57).

Observe that in this as well as in the previous unreliability formula, reliability depends upon two factors: the *S.D. distribution* and *N*. Reliability may be increased by either decreasing the variability or by increasing the number of pupils.

Unreliability of the *Q* in Table 46.—

$$S.D.\ Q = \frac{1.11\ S.D.\ distribution}{\sqrt{2N}}$$

$$S.D.\ Q = \frac{1.11 \times 2.22}{\sqrt{2 \times 24}} = .36$$

We can be practically certain that the true *Q* is between 1.4 $\pm$ 3 (.36).

Unreliability of the ***S.D.*** in Table 46.—

$$S.D._{S.D.} = \frac{S.D.\ distribution}{\sqrt{2N}}$$

$$S.D._{S.D.} = \frac{2.22}{\sqrt{2 \times 24}} = .32$$

We can be practically certain that the true *S.D.* is between 2.22 $\pm$ 3 (.32)

Unreliability of the ***r*** in Table 46.—

$$S.D.\ r = \frac{1 - r^2}{\sqrt{N}}$$

$$S.D.\ r = \frac{1 - (.303)^2}{\sqrt{24}} = .13$$

The true *r* is almost certainly between .303 $\pm$ 3 (.13).

Unreliability of a Difference.—This is one of the most

useful of all the unreliability measures, especially for experimental work where there is an experimental and control group. The *difference* between the means, medians, or other measures for these two groups determines the conclusion from the experiment.

The unreliability of this *difference* determines the value of the conclusion. The following abbreviated formula makes most conclusions slightly conservative.

$$S.D.\ difference = \sqrt{(S.D.\ measure\ \text{I})^2 + (S.D.\ measure\ \text{II})^2}.$$

Suppose that Experimental Group I of 25 pupils were taught by a new method and showed a mean improvement of 18 with an *S.D. distribution of improvements* of 4, while the Control Group II of 36 pupils showed a mean improvement of 16 with an *S.D. distribution of improvements* of 3. The difference in mean improvements, 18 — 16, is 2. Is this difference so reliable that we can be practically certain that if the experiment were repeated upon similar groups, the difference would not become zero or actually favor Group II? Before we can compute the unreliability of a difference it is necessary to compute the unreliability of the measures with whose difference we are concerned. The total process is shown below.

$$S.D.\ mean\ \text{I} = \frac{4}{\sqrt{25}} = .8 \qquad S.D.\ mean\ \text{II} = \frac{3}{\sqrt{36}} = .5$$

$$S.D.\ difference = \sqrt{(.8)^2 + (.5)^2} = .94 +$$

We can be practically sure that the true difference between the two improvement means will be between the obtained difference $2 \pm 3\ (.94)$, i. e., between — .82 and 4.82. Evidently there is some chance that the true difference is zero or even below zero. For the true difference to go below zero would make the experiment favor Group II. The chances are, however, much greater that the true difference is above zero and favorable for Group I and the new teaching method. The way to determine just how much greater the chances are is shown later.

It is possible to compute the unreliability of the difference between *any* two measures. We are far more often concerned with differences between means, but the formula for *S.D. difference* is applicable to the difference between medians, *Q's, S.D.'s, r's,* etc. Just as the formula for the unreliability of a mean was used in order to compute *S.D. mean* I and *S.D. mean* II, preliminary to substituting these measures in the formula for *S.D. difference,* so it is necessary to compute *S.D. r* I and *S.D. r* II or *S.D. Q* I and *S.D. Q* II according to the formula for the unreliability of an *r* and of a *Q* respectively, before substituting in the formula for the unreliability of a difference.

Experimental Coefficients.—Because the unreliability of a difference is so extensively used in experimentation, and because it is difficult for some people to think in terms of chances, I have devised what may be called an *experimental coefficient.* This experimental coefficient is easily computed and automatically shows in a relatively non-technical fashion just how unreliable any difference is. We discovered in computing *S.D. difference* above that the unreliability of the obtained difference of 2 is .94. The experimental coefficient is represented by the following formula:

$$\text{Experimental Coefficient} = \frac{\text{Difference}}{2.78 \quad \textit{S.D. difference}}$$

Since the difference is 2 and the *S.D. difference* is .94,

$$\text{Experimental Coefficient} = \frac{2}{2.78 \times .94} = .76$$

The experimental difference of 2 *is only .76 as large as it needs to be in order that we may be practically certain* that the new method of teaching is truly better than the method used with the control group. Had the difference been 2.61 instead of 2 the experimental coefficient would have been as follows:

$$\text{Experimental Coefficient} = \frac{2.61}{2.78 \times .94} = 1.0$$

An experimental coefficient of 1.0 is just exactly practical certainty. An experimental coefficient of .5 means half certainty, one of 2.0 means double certainty and so on.

If a statement is desired, in terms of chances, of the probability that the true difference is a zero difference or less, i. e., actually favors a conclusion opposite to the one obtained, such a statement in terms of chances, once the experimental coefficient has been computed, may be read directly from Table 47. This table applies to a difference between any two obtained measures as truly as it applies to a difference between obtained means.

TABLE 47

Shows How to Convert an Experimental Coefficient into a Statement of Chances

Experimental Coefficient	Approximate Chances
.1	1.6 to 1
.2	2.5 to 1
.3	3.9 to 1
.4	6.5 to 1
.5	11 to 1
.6	20 to 1
.7	38 to 1
.8	75 to 1
.9	160 to 1
1.0	369 to 1
1.1	930 to 1
1.2	2350 to 1
1.3	6700 to 1
1.4	20000 to 1
1.5	67000 to 1

How to Transmute *S.D. Measure* into *P.E. Measure*. It has already been noted that in a normal frequency distribution, *P.E.*, *M.D.*, and *Q* are equal and that *P.E.* equals .6745 *S.D.* It is somewhat more conventional to express unreliability in terms of *P.E.* than in terms of *S.D.* *P.E. measure* is found by multiplying *S.D. measure* by

.6745. One of the above formulæ is repeated to illustrate this transmutation. It is similar for other formulæ.

$$\textit{S.D. mean} = \frac{\textit{S.D. distribution}}{\sqrt{N}}$$

$$\textit{P.E. mean} = \frac{.6745\ \text{S.D.}\textit{distribution}}{\sqrt{N}}$$

How to Interpret Unreliability. —Illustrations have already shown how to interpret *practical certainty,* which means that the chances are roughly 369 to 1 that the true measure is between the obtained measure ± *3 S. D. measure.*

Other approximate chances are given below:

The chances are 2.15 to 1 the true measure is between obtained measure	± *1 S.D. measure*
The chances are 21 to 1 the true measure is between obtained measure	± *2 S.D. measure*
The chances are 369 to 1 (practical certainty) the true measure is between obtained measure	± *3 S.D. measure*
The chances are 1 to 1 the true measure is between obtained measure	± *1 P.E. measure*
The chances are 4.6 to 1 the true measure is between obtained measure	± *2 P.E. measure*
The chances are 22 to 1 the true measure is between obtained measure	± *3 P.E. measure*
The chances are 142 to 1 the true measure is between obtained measure	± *4 P.E. measure*
The chances are 369 to 1 (practical certainty) the true measure is between obtained measure	± *4.4 P.E. measure*

Summary for Statistical Methods.—Two problems have been used for most of the illustrations in order to reveal the continuous and cumulative nature of statistical processes. To return to the simplicity of childhood, these processes are as continuous and interdependent as "The House That Jack Built." For example, this is a frequency surface, that yields a distribution, that yields a central tendency, that yields a variability, that makes a correlation, etc. As further evidence of this continuity and as a test of the student's mastery of the technique described the following problems have been solved. It is suggested that the student verify the answers.

TABLE 48

Sample Computation of Some Common Statistical Measures.
(Approximate Answers.)

Score	Freq. I	Freq. II	Freq. III	Freq. IV	Score	Freq. I	Freq. II	Freq. III	Freq. IV
0 — 2	3	1	25	1	0 — 10	1	3	1	50
2 — 4	4	1	25	1	10 — 20	1	4	1	50
4 — 6	4	2	50	1	20 — 30	2	4	1	100
6 — 8	5	2	100	0	30 — 40	2	5	0	200
8 — 10	5	4	300	1	40 — 50	4	5	1	600
10 — 12	6	5	200	4	50 — 60	5	6	4	400
12 — 14	4	4	100	2	60 — 70	4	4	2	200
14 — 16	4	3	50	0	70 — 80	3	4	0	100
16 — 18	3	2	25	0	80 — 90	2	3	0	50
18 — 20	2	1	25	2	90 — 100	1	2	2	50
Mode	11.00	11.00	9.00	11.00		55.00	55.00	55.00	45.00
Mean	9.55	10.76	9.89	10.50		53.80	47.75	52.50	47.45
Median	9.60	11.00	9.67	11.00		55.00	48.00	55.00	48.35
Q_1	5.50	8.13	8.17	7.00		40.65	27.50	35.00	40.85
Q_3	13.50	13.88	11.75	13.00		69.40	67.50	65.00	58.75
Q	4.00	2.88	1.79	3.00		14.40	20.00	15.00	8.95
S.D. from Mean	5.08	4.39	3.54	5.30		21.95	25.40	26.50	1.77
S.D. mean	.80	.88	.12	1.53		4.40	4.00	7.65	.04
P. E.	.43	.46	.06	.81		2.30	2.15	4.05	.02

SUPPLEMENTARY READING FOR PART III

ALEXANDER, CARTER.—*School Statistics and Publicity;* Silver, Burdett & Company, New York, 1919.

BRINTON, WILLARD C.—*Graphic Methods for Presenting Facts;* The Engineering Magazine Company, New York, 1917.

KING, W. I.—*Elements of Statistical Methods;* The Macmillan Company, New York, 1912.

ROUTZAHN, E. G., and MARY S.—*The A B C of Exhibit Planning;* Russell Sage Foundation, New York, 1918.

RUGG, HAROLD O.—*Application of Statistical Methods to Education;* Houghton Mifflin Company, New York, 1916.

SCOTT, WALTER DILL.—*The Psychology of Advertising;* Small, Maynard & Company, Boston, 1910.

THORNDIKE, EDWARD L.—*An Introduction to the Theory of Mental and Social Measurements;* Teachers College, Columbia University, New York, 1913.

APPENDIX

HOW TO SECURE TESTS AND DIRECTIONS FOR THEIR USE

This book has attempted to give the fundamental procedure for any type of mental measurement. But it has been impossible in a book of this size and inappropriate in a book of this character to describe at length the existing tests and scales. Tests are changing at a phenomenal rate and changing for the better. It is the function of frequent bulletins issued by book companies and bureaus of research to inform educators of the latest and best tests. All the important centers which distribute testing material are prepared to send free or practically free literature describing their tests. More than this they are glad to give expert advice as to the test or tests which it is best to use in a particular situation. Finally, they are prepared, for a small charge, to send for inspection sample tests. Again, the bureaus which issue tests usually do and always should send with the tests which have been ordered a leaflet giving detailed directions for applying and scoring the tests, for tabulating results, and for computing pupil and class scores. The directions usually include norms for the test and frequently suggestions for the uses of results. As a precaution the individual, when writing for tests, should request that all necessary directions for properly using them be sent.

The following are the chief centers for the distribution of tests:

Bureau of Publications, Teachers College, New York City.
Public School Publishing Co., Bloomington, Ill.
World Book Company, Yonkers-on-Hudson, N. Y.
C. H. Stoelting Company, Chicago, Ill.

The World Book Company has just issued a booklet entitled *Bibliography of Tests for Use in Schools,* which sells for ten cents. This booklet gives tests sold by other agencies than themselves. To date this is probably the most complete list of tests ever assembled. The other centers mentioned above also have descriptive booklets for the tests which they distribute.

The following references also contain elaborate lists of tests or bibliographies on tests or both:

Holmes, Henry W., and Others.—*A Descriptive Bibliography of Measurement in Elementary Subjects;* Harvard University Press, Cambridge, Mass., 1917.

National Society for the Study of Education.—*Seventeenth Year Book, Part II;* Public School Publishing Company, Bloomington, Ill., 1918.

Ruger, Georgie J.—*Bibliography on Psychological Tests;* Bureau of Educational Experiments, New York, 1918.

Whipple, Guy M.—*Manual of Mental and Physical Tests, Vols. I and II;* Warwick & York, Baltimore, 1910.

INDEX

Date Due

Je 7 '48			
~~FACULTY~~			
OCT 27 1995			
NOV 20 1995			